LEGAL SERVICE PROVIDER SCORING SYSTEM

STUDY REPORT

Author: Mr Howard Chen

Contributor: Dr Dewei Yi

Editor: Mr Fraser Matcham

1st September 2021

Funded

THE UNIVERSITY of EDINBURGH

In collaboration

1495 UNIVERSITY OF ABERDEEN

ABSTRACT

This project was commissioned, on application, by Legal Utopia and generously funded by The University of Edinburgh with academic collaboration with the School of Natural and Computing Sciences, The University of Aberdeen.

The aims of the project were to calculate the weights of quality indicators (QIs) designated by Legal Utopia for scoring and evaluating all SRA-regulated legal services providers (LSPs) in the UK. This includes those designated as a recognised body, licensed body, or recognised sole practice, including those with the same designator with conditions.

The SRA-regulated LSPs included two datasets of different dates during the 6-month review Legal Utopia undertook, of which, 3-months included the machine learning modelling. The first dataset was representative of the February 2021 SRA-register data and the second dataset was representative of the July 2021 SRA-register data.

These datasets, respectively, were organically merged with historical user data from Legal Utopia's 'Find-A-Lawyer' service on users' searching behaviour, as well as all SRA-regulated LSPs' characteristics. A special and detailed data pre-processing exercise was carried out to make up the unselected SRA-regulated LSP samples which were not recorded in the Find-A-Lawyer service or Legal Utopia's mobile app.

Both classical machine learning and deep learning neural network algorithms are implemented for recognising the patterns of selected and unselected LSPs. The hyper-parameters of each machine learning model are turned to suite each specific dataset that they are trained on.

The full dataset, representing the merged dataset referred to above, was severely imbalanced. This was due to the high number of LSPs compared to the relatively low number of contacted LSPs from the users of the Find-A-Lawyer service. To resolve this issue, both cost-sensitive learning algorithms and cost-sensitive learning resampling methods are employed to improve machine learning models' performance. After cross-comparison, the "balanced dataset" is identified as the most suitable candidate for the feature importance analysis.

There were three machine learning interpretation methods implemented to calculate the feature importance. Through a comprehensive evaluation and literature review, the SHAP method is deemed the superior approach over the native and permutation feature attribution methods.

The SHAP method is finally applied to the best performed Gradient boosting model to calculate the QI weights (feature importance). The interpretation results are explained at both the global and local level at the end of this study.

SPECIAL THANKS

Legal Utopia would like to bring particular thanks to Howard Chen – Machine Learning Engineer – for his diligent and hard-work since joining the team in May 2021 to focus his time on conducting this data review and machine learning research, as well as taking the time to draft this study alongside his MSc in Artificial Intelligence dissertation.

Howard Chen

Machine Learning

Engineer

Legal Utopia also thanks Intisar-al Haque – Software Developer - for his work on the regional branch cost modelling and geo-encoding work to support this project, as well as the paralegal and legal research team members that undertook the manual review of the SRA data and data standardisation work on practice area assignment.

Legal Utopia also recognises that this research and development would not be possible without the funding support of The Data Lab at The School of Informatics, The University of Edinburgh, as well as the collaborative engagement and input of Dr Yi of The University of Aberdeen.

Executive Summary

This, University of Edinburgh funded, study into a Legal Services Provider Scoring System undertaken by Legal Utopia in collaboration with The University of Aberdeen is an innovative study comprising of an accumulated 6-month pre-requisite stakeholder engagement exercise, as well as a 6-month data review and parallel 3-month modelling exercise between 2020/21. This study comprises the findings of these research and development exercises, as well as presenting the commercial impacts to Legal Utopia's services.

The key summary findings of this research study are:

- The SRA maintains a database of 23,434 entities, comprising the total number of rows of information accessible from the SRA API, of which, in February 2021, only 10,003 were actively regulated and authorised by the SRA;
- The SRA also maintains branch/office-based data on the 23,434 entities on their register, this includes, as of February 2021, 30,849 rows of information accessible from the SRA API;
- The SRA entities that are actively regulated and authorised by the SRA reduced at a rate of 22.5 firms per month between February 2021 and July 2021 from 10,003 to 9,913;
- The SRA entities that are actively regulated and authorised by the SRA had almost 3,500 entities with 10-years authorisation history (number of years authorised by the SRA) as of February 2021, but this reduced by also 300 firms by July 2021;
- The number of new actively regulated and authorised entities by the SRA has increased from fewer than 100 firms to almost 300 between February 2021 and July 2021;
- The actively regulated and authorised entities by the SRA have an unbalanced authorisation type with approximately 7,000 firms authorised as a 'Recognised Body' (with or without conditions) but only 2,000 as 'Recognised Sole Practice' (with or without conditions), and 1,000 as a 'Licensed Body'. This went relatively unchanged between February 2021 and July 2021;
- The data held by the SRA of all entities on the Register had approximately 350 entities with same trading names duplicated across the Register with just over 300

having 2 duplicates associated with multiple SRA entities and a maximum number of duplicates associated with multiple SRA entities being 19;

- The data held by the SRA, as of February 2021, showed that approximately 700 SRA entities has 2 branches/office locations with the highest number being 11 offices to a single entity, however, by July 2021 this went largely unchanged other than the prevalence of those with 11 offices reduced negligibly. Of all firms with more than 2 offices, none had 8 or 10 offices but the maximum number of offices to one entity was 11;
- The data held by the SRA and produced by its API had no entity's geo-location data, no Google Business ID, no website security or status ID, no trading history value, no pricing value, no turnover values, no cybersecurity or insurance credentials, no regulatory history, and no SRA badge compliance status. There was no attribution of individual solicitors or lawyers that worked for or with an SRA entity;
- The data held by the SRA and produced by its API had no standardisation of the "Work Area", "Freelance Basis", or "Reserved Activities" data points. All data points recorded on the SRA Register API data were only mandated to be updated annually and only by actively regulated and authorised entities;
- The data held by the SRA of all entities, only 52% had a URL website link available within the dataset and 48% without, this does not however factor website quality, security, loading speed, design, or working status of which, on manual review, scored poorly in all fields;
- The study sought to quantify the importance of 10 quality indicators (Authorisation Duration, Cybersecurity Credentials, No. of Practice Fields, Customer Reviews, Cost, Location, Regulatory History, Website, No. Payment Options, and Insurance Coverage) identified by Legal Utopia, however, due to the lack of regulator engagement and poor data landscape, the study was conducted against only four quality indicators;
- The study focused on the following as "quality indicators": Authorisation Duration, Number of Practice Field(s), Location, and Website Availability.
- The study sought to classify SRA entities using machine learning to determine if one firm over another could be predictably "comparable" based on historical user data searching behaviour.
- The historical user data of contacted SRA entities, via Legal Utopia's Find-A-Lawyer service, had 951 samples with attributed data for the purposes of machine learning modelling;

- The modelling was applied across 8 models, including a multilayer perception neural network, and achieved an accuracy on average of 96%, however, failed in all cases to identify a contacted class.
- The modelling with a balanced dataset achieved an accuracy on average of 73% across all 8 models used;
- The modelling with a synthetic, balanced dataset achieved an accuracy on average of 68% across all 8 models used;
- The modelling with additional features achieved an accuracy on average of 74% on a SMOTE dataset across all 8 models used;
- On interpretation of all the modelling results, the balanced dataset was deemed most suitable of 6 modelling exercises for the study's purposes at a 73% accuracy overall using a Gradient Boosting Model;
- The study sought to use machine learning and historical user data, along with the four identified "quality indicators" applied to all SRA entities, to determine the computed "importance weighting" of each of the four QIs as features that influence users when choosing to contact an LSP;
- On interpretation of all modelling results of QI importance, the majority of models determined the number of authorisation years of an LSP is the most influential with office location in the same area as the user being the least important contributing feature of the four QIs. The website availability was third and no. of practice fields (practice area coverage) being second, as well as most controversial as it scored first at least once across all rankings in all models;
- The QI importance weights were modelled on balanced data and across 8 models with an average overall accuracy of 73% and an AUC of 78%;
- The study found that, on existing data, the modelling results are unsatisfactory due to a severely imbalanced dataset and, as a result, a prototype of the modelling is not appropriate for user testing or market release; however, the results from the SRA data insights indicate a reduction in the overall number of actively authorised and regulated entities. This combined with an increasing dataset of users' searching behaviour and with optimisation is likely to make a data-driven comparison algorithm and QI feature importance computation authoritatively possible for user testing and market release in the near future.

Table of Contents

1. INTRODUCTION

1.1. Commercial Background and Problem Statement

Legal Utopia has developed a mobile app to provide this service for its customers within the UK. One of the main functionalities of this mobile app is to help users find a legal services provider in accordance with the users' preferred location and required legal practice field(s). This mobile app lists all relevant candidates on a Google Map interface for the users' selection.[1]

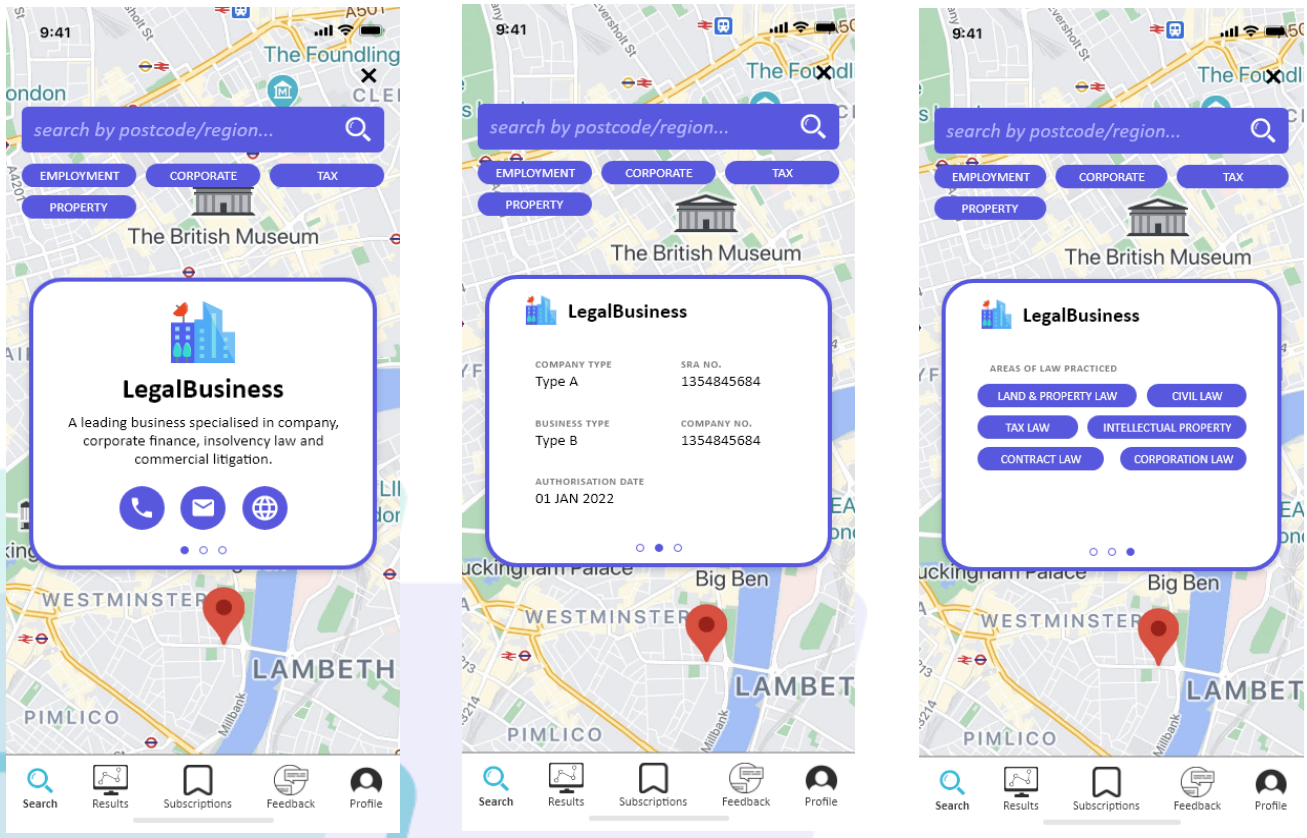


Figure 1

[1] See Figure 1 – Find-A-Lawyer graphical representation (June 2021)

The mobile app displays all matched LSPs on the screen as red drop-pins or blue circular groupings representing several red drop-pins in a concentrated location. For most of searching cases, a large number of red balloons are displayed on the screen without any indication of their quality or characteristics, which leads to an inefficient searching experience (Problem Statement).

Legal Utopia, therefore, plans to develop an evaluation system to score all LSPs registered in the UK against the designated QIs. In this way, only the highest scoring LSPs will be displayed or recommended to users, which improves users' selection efficiency.

Such a scoring system can be designated by professionals in the legal field, who have the domain knowledge of designated QIs. However, Legal Utopia determined, in this project, to develop a "data-drive" evaluation system to understand what characteristics or features of LSPs influence the users' selection. In other words, the QIs importance knowledge will be learnt from the user data rather than from field experts. This data-based approach provides a different insight and is believed to be a favourable supplementation to the traditional evaluation system.

1.2. Task Definition from the Data Science Perspective

Legal Utopia has collected all users' searching history, including the date and time a user has used the app, the location, and the practice field the user has searched, and finally the LSP that the user has selected/contacted. In addition, Legal Utopia has collected various spreadsheets containing LSPs QIs and logistic information.

In order to develop a data-drive scoring system required by Legal Utopia, it is proposed to use machine learning algorithms to find the commonalities among the LSPs which have been selected by users. In other words, it is expected to find out the characteristics and features of the LSPs which led the users to select them over the others.

In addition, the influence of each QI will be quantified and converted to weights. Applying weights to each LSP corresponding QIs will obtain the overall scores of the LSPs. All LSPs shall then be ranked as per their score[2] and the highest-scored LSP will be recommended to customers.[3]

[2] See Figure 2 – Find-A-Lawyer (Concept) ranking graphical representation

[3] See Figure 3 – Find-A-Lawyer (Concept) LSP score graphical representation.

1.3. Proposed Solutions

Based on the problem definition, the project scope can be divided into the following tasks:

Task 1: Generate a single dataset which represents users' selection behaviour. This includes necessary data merging, cleaning, wrangling, and scaling work.

Task 2: Build and train machine learning models to predict users' selection of LSPs. Since the mobile app data records the LSPs that have been selected by users, this is a supervised classification problem. The following eight commonly used classification machine learning models will be employed for fulfilling this project's purpose:

- Logistic regression
- Random Forest
- Gradient Boosting
- SVM
- KNN
- Deep Neural

Task 3: Turn the hyper-parameters of each machine learning model for each specific dataset to achieve a highest possible accuracy.

Task 4: Apply appropriate techniques/methods to handle the imbalanced data. Evaluate their performance and select the best candidate.

Task 5: Apply appropriate machine learning interpretation techniques/methods to compute the QI weights (feature importance) based on machine learning model prediction results. Compare the interpretation of methods' performance and select the most suitable approach for finalising the feature importance.

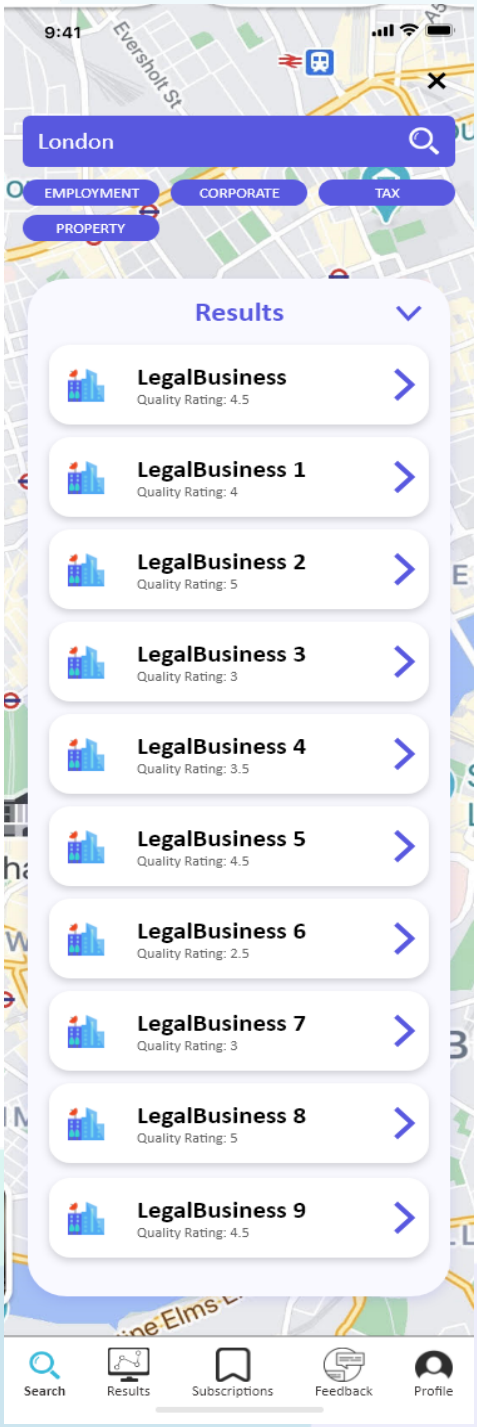


Figure 2

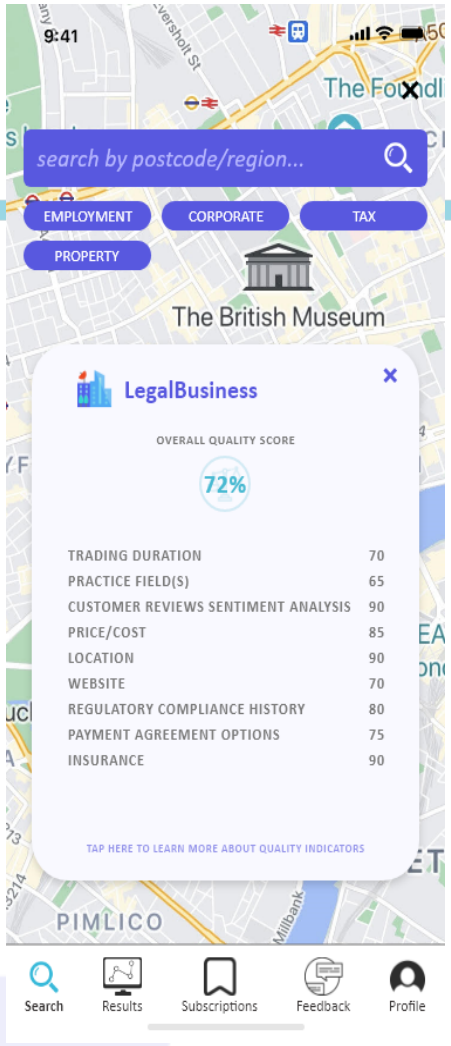


Figure 3

2. DATASET PROPOSAL

2.1. Data Overview

2.1.1. Data Source Selection

After checking all datasets provided by Legal Utopia, the following two datasets are deemed the most relevant to this project:

1. Mobile app user data; and
2. Solicitors Regulation Authority (SRA) Register spreadsheets.

Note: New mobile app user data is collected continuously, and the SRA data is updated on a regular basis.

2.1.2 Mobile App User Data

At the time this project is commenced, Legal Utopia exports the mobile app user data in a format of CSV files. It contains 1140 rows in total, each row represents a user searching event.

The mobile app user data contains a large number of columns, which records comprehensive information about the users. The details are listed below:

- User ID
- Firm: legal service provider that has been contacted by user

- Search: the location of legal service provider that the user is looking for
- Date
- Account type
- Event count: How many times the user has searched for a legal service provider
- Latitude and Longitude
- City ID
- Country
- Region
- Continent and Continent ID
- Language
- App store and App version
- Browser and Browser version
- Operating system and OS version
- Device, Device model and Device category
- App screen name, Screen resolution, Screen name and Screen class
- Page title
- Active users

2.1.3. Solicitors Regulation Authority Data

In addition to the mobile app user data, Legal Utopia has collected the business information of all LSPs registered in the UK from the SRA. Due to the way the SRA organises information, the data is provided in two CSV files, namely "SRA Organisation" and "SRA Offices". It is not possible to export all information in a single file, which brings challenges to the data processing work, which will be discussed in section 3.4 in detail.

The "SRA Organisations" sheet identified each Legal Utopia firm as an entity and assigns a unique ID number. There are 23434 rows in total, each row records the information of a unique LSP firm. The columns contained in the "SRA Organisations" sheet are listed below:

- (Organisation) ID
- SRA Number
- Practice Name
- Authorisation Type
- Authorisation Status
- Authorisation Date
- Organisation Type
- Authorisation Status Date

- Freelance Basis
- Regulator
- Trading Names
- Previous Names
- Work Area
- Websites
- Reserved Activities
- Company Reg No.
- Constitution
- No. Of Offices
- Type
- CH Company Type
- Office Other

It should be noted that the data collected by the SRA did not include essential data for geo-mapping LSPs and, as such, Legal Utopia undertook a geo-encoding exercise to identify the longitude and latitude of each office and branch of every SRA entity. In undertaking this exercise, it was identified that 405 addresses supplied by the SRA were not recognised by Google Maps, this was either due to incomplete address data, no address data being available, or, in limited cases, the addresses provided not existing.

The data also did not have adequate coverage of whether the SRA entities had active websites or the website URL address, as well as whether the website had SSL encryption. The data also had most of the email and phone number data missing and failed to collect comprehensive or accurate information on SRA entity company numbers and structures. The vast majority of company structure data was missing despite a company number being available. In a limited number of cases, company numbers or company structures were incorrectly labelled.

Legal Utopia subsequently undertook semi-automated review of the company numbers and entity names using the Companies House API to correctly determine all company structures and quickly identify incorrect or invalid company numbers assigned to SRA entities. Although the majority of this exercise was undertaken automatically, it was not 100%, and, as such, Legal Utopia undertook a manual review in a small number of cases.

Legal Utopia also had to undertake a manual review exercise of the emails and phone numbers missing from being assigned to entities on the "SRA Organisations" data. Therefore, Legal Utopia supplemented the data on the SRA Organisations spreadsheet with the following:

- Websites
- Active/Inactive
- SSL Encryption
- Longitude/Latitude

- Email
- Phone
- Company Structure
- Company Number

In contrast to the "SRA Organisations" sheet, the "SRA Offices" sheet identifies each office of a LSP as an entity and assigns a unique office ID number accordingly. Not surprisingly, the total number of rows is higher, reaching 30849. Each row records the information about a unique LSP office. The columns contained in the "SRA Offices" sheet are listed below:

- (Organisation) ID
- Office ID
- (Firm) Name
- Address
- Postcode
- Town
- County
- Country
- Phone Number
- Website
- Email

- Office Type

The data from the SRA Offices spreadsheet was also missing the longitude and latitude of the offices and branches, which are of a greater number than the total number of SRA-regulated entities, as well as the Google Place ID of those offices. As such, Legal Utopia had to undertake another geo-encoding exercise to generate and supplement the SRA data with the following:

- Google Place ID
- Latitude
- Longitude

However, it is noted that this spreadsheet tended to be more comprehensive in the data collected with the vast majority of address information available, as well as phone numbers and emails, including an office type accurately assigned to assist with distinguishing between a headquarter and an office branch.

2.2. Dataset Proposal

As proposed in 1.3, classification machine learning models shall be used to analyse users' selection patterns. A suitable dataset, containing relevant features and targets are required for representing users' selection behaviour.

2.2.1. Target of the Dataset

The "Firm" information contained in the mobile app user data is selected as the target of the dataset, as the "Firm" column identifies which LSPs have been selected/contacted by users.

2.2.2. Features of the Dataset

The features of the data set shall meet two criteria:

Firstly, they will cover the QIs identified by Legal Utopia for quantifying a LSP's quality, including:

- Trading Duration
- Cybersecurity Credentials
- Practice Field(s)
- Customer Reviews Sentiment Analysis
- Price / Cost
- Location
- Regulatory Compliance History
- Website
- No. Payment Agreement Options
- Insurance (Cyber & Data, PII)

Due to the lack of regulator engagement and data availability of QI information at the time of conducting this project, Legal Utopia has narrowed down the primary QIs to the following four items:

- Trading Duration
- Practice Field(s)
- Location
- Website (availability)

Secondly, these features should be displayed on the mobile app user interface, which means they are disclosed to the users and were taken into consideration when the users were selecting LSPs.

The following items are displayed on the current mobile app when a user opens the Find-A-Lawyer service and clicks on a LSP's icon, including:

- Phone contact
- Email link
- Website link
- Company type
- Business Type
- Authorisation Date
- SRA No.

- Company No.
- Areas of Law Practiced

2.2.3. Dataset Structure Proposed

Combining the target and features requirements in Section 2.2.1 and 2.2.2, the overall dataset structure is proposed as below:

	Features	Information Source	Consideration
Features	Legal service provider's authorisation duration.	SRA Organisations Sheet.	This indicates how many years this firm has been authorised as a legal service provider. A longer service history may be appreciated by customers.
	Total number of the legal service provider's practice fields.		This indicates how comprehensive this firm's legal business issue, and indirectly imply the scale of the firm.
	Website availability of the legal service provider.	SRA Organisations Sheet.	This indicates the professionalism and potentially the service quality of a legal service provider.
	If the legal service provider is in the same town as the user's searching town.	SRA Offices Sheet	Customers may prefer to select a legal service provider from the same area as they are based.
Target	The name of legal service providers that have been contacted.	Mobile app user data	This distinguishes the legal service providers that have been contacted versus those that have not.

Table 1

Clearly, there is a need for merging the "SRA Organisations" dataset and the "SRA Offices" to provide the full feature information.

Note: The data merging operation does not refer to simple dataset combinations, but organic merging, the operation details will be covered in section 3.

2.2. Defect of the Provided Dataset

The current mobile app only records the LSPs that have been contacted by users, lacking the records of the LSPs that have been browsed but not contacted. Ideally, the mobile app user data is expected to record the full user behaviour, so that the machine learning algorithms will learn the commonalities of the selected LSPs as well as that of the unselected LSPs.

Unfortunately, the unselected LSPs names are not available. The proposed solution is to remove the selected LSPs instances (1044) from the SRA dataset (23434) and treat the remaining ones as unselected LSPs. This approximation is deemed a reasonable remedy as the SRA dataset is the informed source for the mobile app, users have access to all the LSPs in the SRA dataset.

This proposal requires an organic merge of the mobile app user dataset with the combined SRA dataset, the operation details will be covered in section 3.

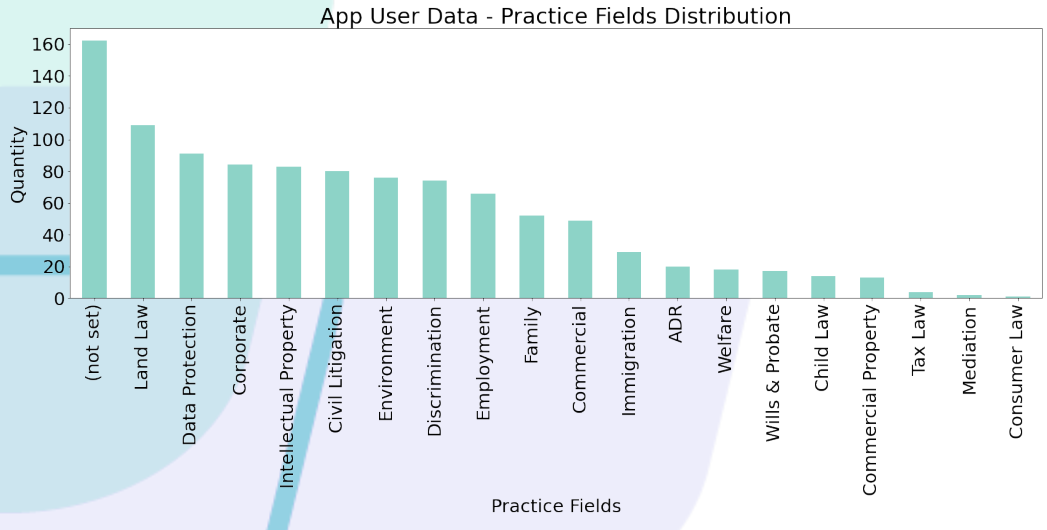
3. DATA PROCESSING

Following the dataset proposal above, this section aims to demonstrate the actions required for generating such a dataset, including necessary data cleaning, data wrangling, and data merging operations.

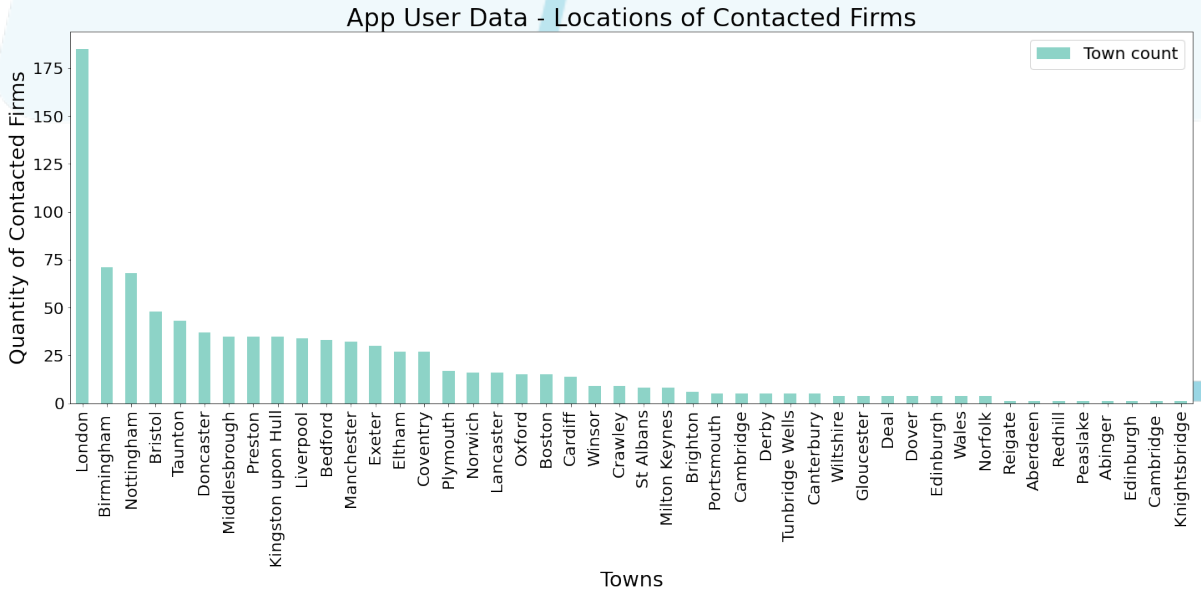
3.1. Processing the mobile app user data

Necessary cleaning work on the mobile app user data is carried out. A critical operation is to delete the 96 instance which do not record any contacted LSP. They occupy a small portion as compared with the total 1140 instances. The rest of the section explores the app user data from various perspectives:

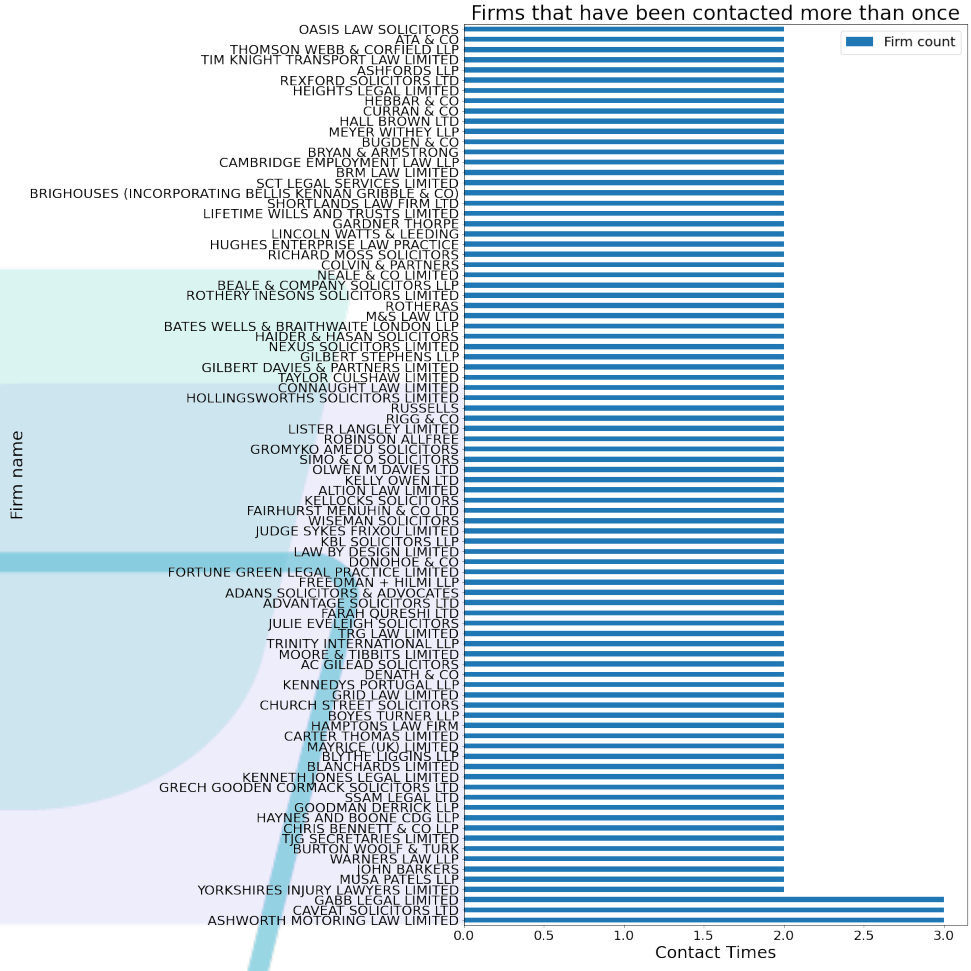
The graph below depicts the practice fields that users have searched for:



The graph below demonstrates the towns from where users have searched for LSPs:



The graph below counts the times the selected LSPs have been contacted for:

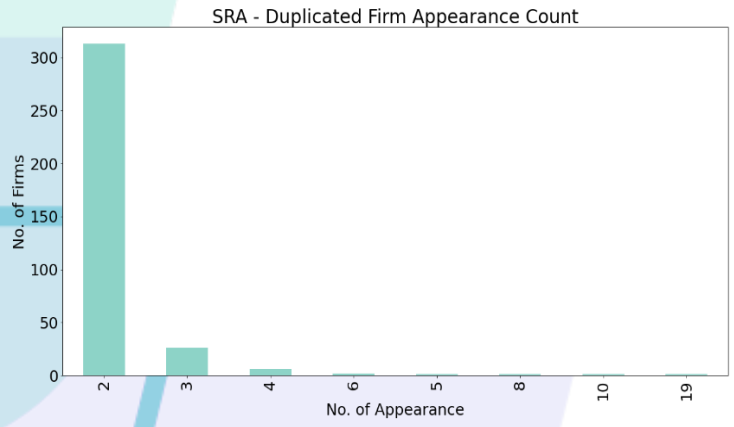


The majority have been contacted once, 85 firms have been contacted twice, and 3 firms have been selected 3x times, which is the highest count.

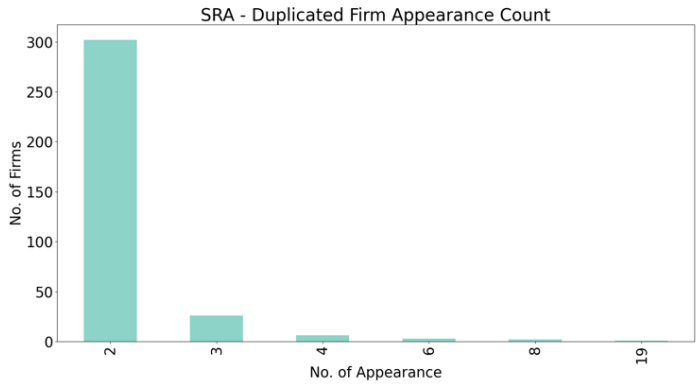
3.2. Processing the SRA Organisations' Data

The SRA Organisations' data processing includes trimming un-needed columns and renaming remaining columns. An important operation is to check the LSP assigned with different ID numbers but sharing the same trading names. This is caused by the branches located in different towns.

The graph below shows the appearance count of the SRA entities contained in the SRA Organisations dataset that have a duplicate trading name and the number of times that trading name appears within the dataset. This chart implies that there are around 350 entities with duplicate trading names and for the vast majority (310) these appear twice with the maximum appearance rate being 19.

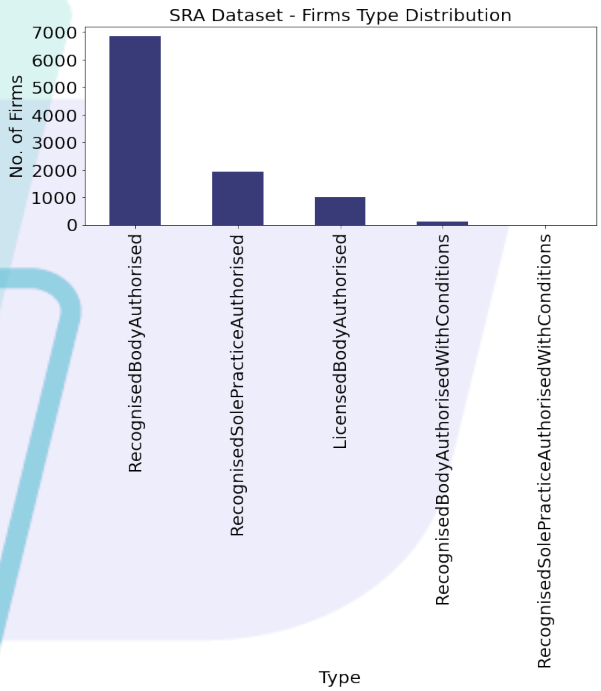


The graph below shows the same as identified above but is based on the July 2021 SRA data and as expected, shows limited difference in trading name duplication.

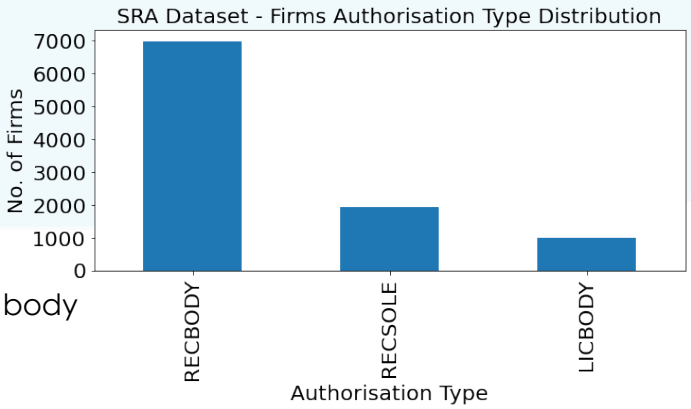


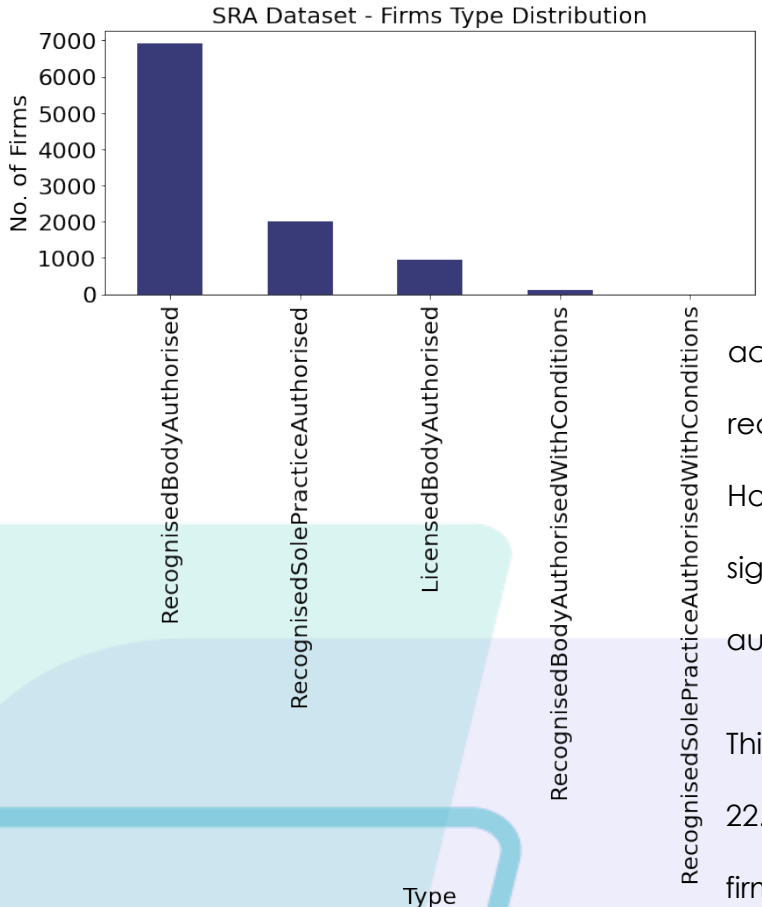
The chart below depicts the distribution of regulated SRA firms distinguished by authorisation type as of February 2021. It shows 10,003 firms actively regulated by the SRA.

Of the 10,003 firms, the vast majority (7,000) were classed as recognised body authorised with recognised sole practice authorised being the second most popular (2,000). This is followed by licensed bodies (1,000).



The chart to the right depicts the distribution of the SRA authorised firms as of February 2021. It shows the 10,003 firms by recognised or licensed body type. This is consistent with the above considering the aggregated recognised body and recognised sole practice.

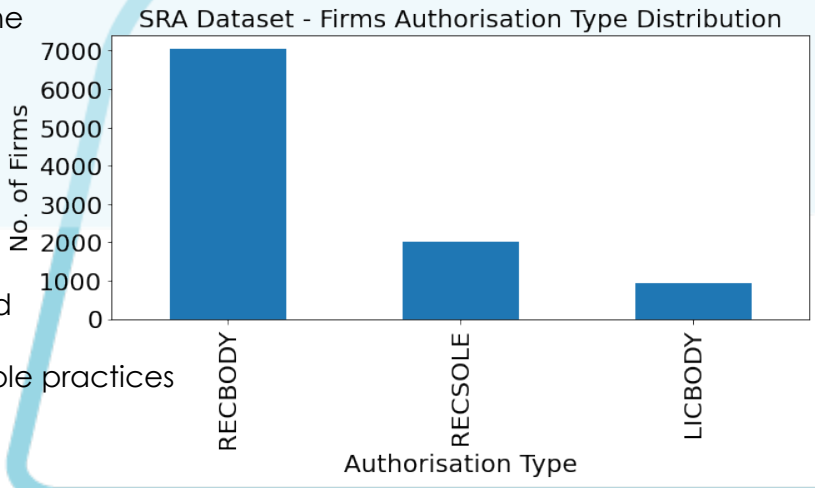




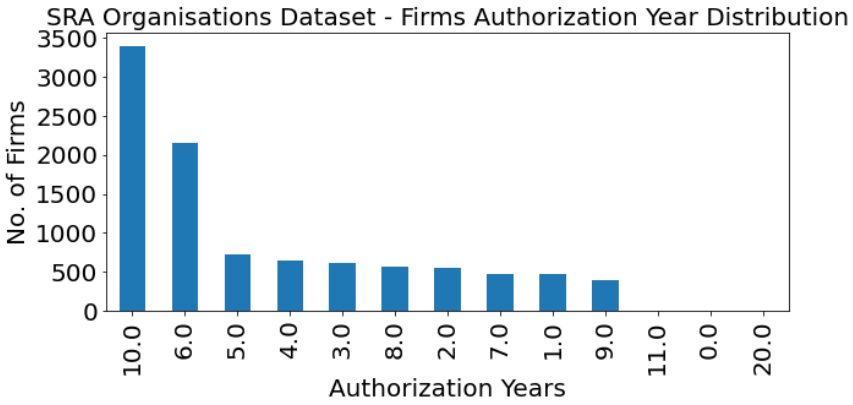
The chart to the left depicts the distribution of regulated SRA firms distinguished by authorisation type as of July 2021. It shows 9,913 firms actively regulated by the SRA, a reduction over four months of 90x firms. However, the reduce does not present a significant bias to any one type of authorised firm.

This chart implies a monthly reduction of 22.5, previously actively regulated, SRA firms between February 2021 and July 2021.

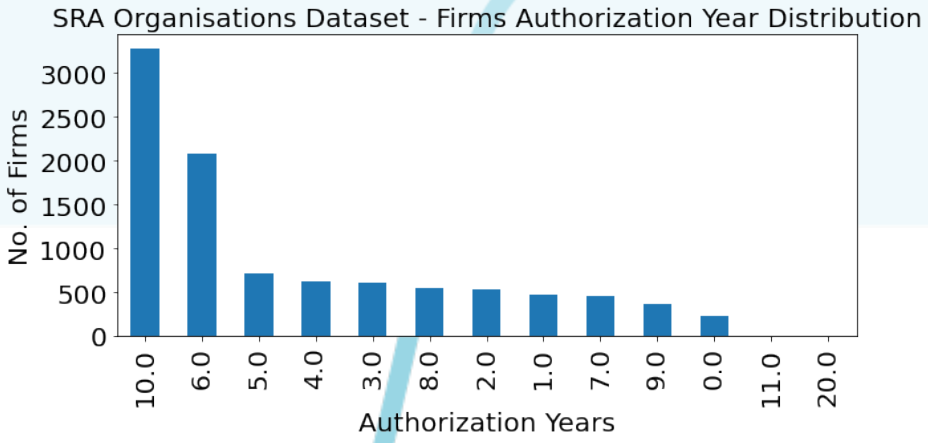
The chart to the right depicts the distribution of the SRA authorisation firms as of July 2021. This is consistent with the above chart considering the aggregated recognised body and recognised sole practices



.



The chart above depicts the number of accumulated authorisation years of the actively regulated entities as of February 2021. It implies the vast majority of the 10,003 firms had been authorised for 10-years (3,500) with a further 2,000 for a 6-year period. However, the number of firms between 0 and 5 are evenly distributed by number, indicating a very low number of new entities seeking SRA-authorisation. This would support the implied finding of a reduction, month-on-month, of actively regulated SRA firms.



The chart above depicts the same accumulation of authorisation years, but as of July 2021. This indicates a significant reduction in the number of firms with 10-years authorisation.

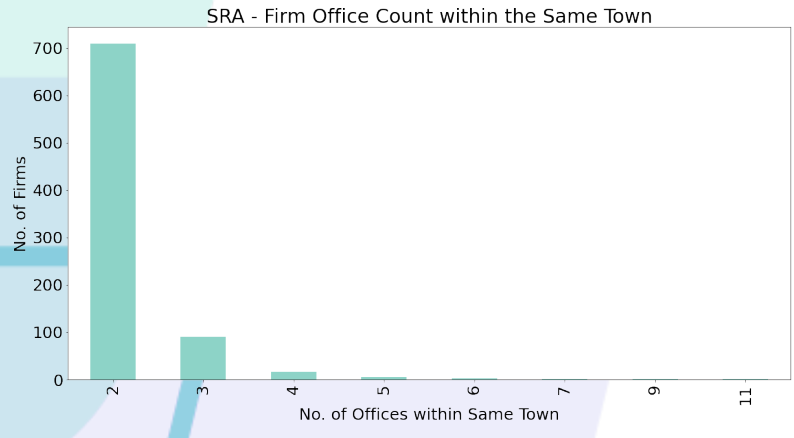
In the course of a four-month period, approximately 300 firms disappeared from this metric and implies that, between February 2021 and July 2021, those actively regulated firms that closed (22.5 per month) were firms with a higher cumulative number of authorisation years than those that did not. Meanwhile, the number of firms with beyond 10-years accumulated authorisation years did not increase or decrease at all.

An additional finding is the number of firms with 0-years of SRA authorisation increased by almost 300. This data suggests more senior firms are closing in Q3 2021 and more new SRA firms are being authorisation at the same time.

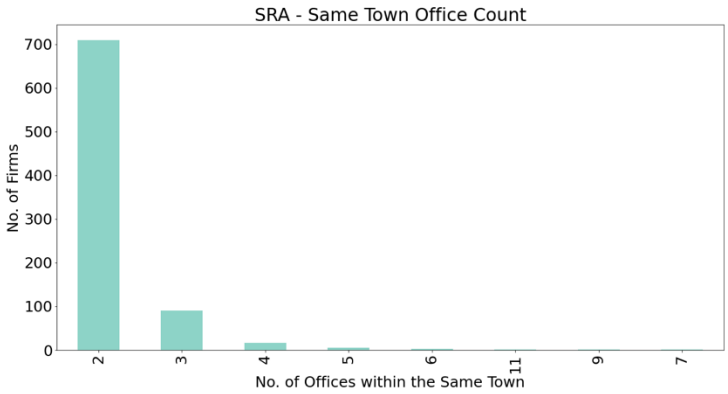
3.3. Processing the SRA Offices Data

The basic SRA Office data processing is similar to that of the SRA Organisations' data. A critical operation is to check the LSP's presence within the same town. As pointed out in section 2.1.3, the "SRA Offices" dataset assigns a unique ID to each office of an LSP. The problem is that the mobile app only records the LSP name that has been selected rather than the specific office associated. Before merging the "SRA Organisations" sheet with the "SRA Offices" sheet, it is necessary to remove the multiple offices of a LSP within the same town. Only one office from each LSP will be retained.

The graph below shows, as of February 2021, the office count of SRA entities (around 750) with more than 1 office or branch, with the vast majority having 2x offices or branches with the maximum, in a couple instances, being 11x offices or branches.



The graph below shows, as of July 2021, the same office count as identified above and as expected, shows no difference in the number of offices.



3.4 Merging the "SRA Organisations" Data with the "SRA Offices" data

Section 2.2 has discussed the necessity of merging the "SRA Organisations" dataset with the "SRA Offices" dataset.

3.4.1 Merging Operation

The first step is to change the firm and town names to lower case as capital and lower-case names are inconsistently used in different data sheets. This operation ensures that the same firm names are aligned regardless of their original letter case.

Both datasets are imported as Pandas data frames. Pandas provides different options for merging data frames.[4] For this case, the "ID" column, a unique identification of a LSP is set as the merging key. Two datasets will be aligned as per ID numbers. The join method is set to be "inner", so that the intersection of the two data frames are retained. The code snippet is shown below:

```
SRA=pd.merge(SRA_1,SRA_2_Rem_dupic_town,how='inner', on='Id')
```

3.4.2. Duplication Check and Removal

Duplication cannot be fully eliminated due to the different way the two datasets organise information (unique firm identification and unique office identification) and the native errors of the two datasets. Duplication check is always necessary after the merging operation.

It is discovered that 477 LSPs have multi-presence in the same town. For example, the firm "Westbrook Law Ltd" is counted three times with different postcodes in London. It is likely this firm has three different offices in London. Similar duplicated offices like this example should be deleted due to the reason stated in Section 3.3.

[4] Merge, join, concatenate and compare.

3.5 Merging the concatenated SRA Data with the mobile app user data

The next step is to merge the concatenated SRA data with the mobile app user data.

3.5.1. Merging Operation

The first step is to change the firm and town names to lower case as capital and lower-case names are inconsistently used in different data sheets. This operation ensures that the same firm names are aligned regardless of their original letter case.

The "Firm" column is set as the merging key, it is a unique identification in both the SRA data and the mobile app user data. The code snippet is shown below:

```
df_firm_match=pd.merge(SRA_Rem_dupic_town,App_df,how='left', on=['Firm'])
df_firm_match
```

3.5.2. Duplication Check and Removal

For the same reason as stated in Section 3.4.2 duplication check is carried out. It is discovered that 291 LSPs have multi-presence in the same town. For example, the firm "Morgan has solicitors Ltd" has two offices in London, assigned with different postcodes. Similar duplicated offices like this example should be deleted due to the reason stated in Section 3.3.

3.6. Dataset Feature Engineering and Target Labelling

After merging all required dataset, the next step is to generate the features and target columns to implement the proposal raised in Section 2.2.3.

3.6.1 Generating the Feature Columns

Feature Id	Feature Description	Operation
1	Legal service provider's Trading Duration	This is to calculate each legal service provider's authorisation duration. The SRA dataset only provides the authorised date, hence the duration is calculated from the authorised date to the project start date
2	Total Number of each Legal Service Provider's Practice Fields	This is to calculate the total number of each legal service provider's practice fields. The SRA dataset provides the names of all practice fields in one row, the operations are segmenting the name-string and count the quantity of individual practice fields
3	Websites Availability	The mobile app user interface is currently designed to be location based, which means the mobile app always navigate to the location that the user searched and display the legal service provider in that area.
4	If the Legal service provider is in the same town that the user searched for	The first step is to list all towns in where the legal service providers have been selected by all users. The second step is to identify the legal service provider from the SRA dataset which are located in the towns as listed in step 1. The third step is to create a "Same Town" column, which is feature 4, the legal service providers identified in step 3 are marked "yes", the remaining are marked "No"

Table 2

3.6.2. Generating the Target Column

The Target Column contain all LSP names in the SRA dataset. The ones that have been selected by the mobile app users are marketed "yes", the remaining ones are marked "no".

3.7. Dataset Check and Tidy-Up

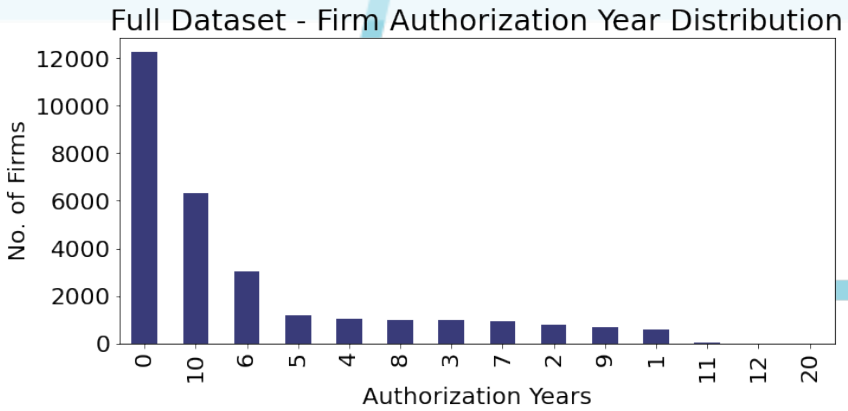
The dataset proposed in Section 2.2 is almost completed at this step. Data check and tidy up operations are carried out as listed below:

- Check the data type of each column
- Check for missing values in the entire dataset and fill the empty cells
- Encoding all categorical feature columns

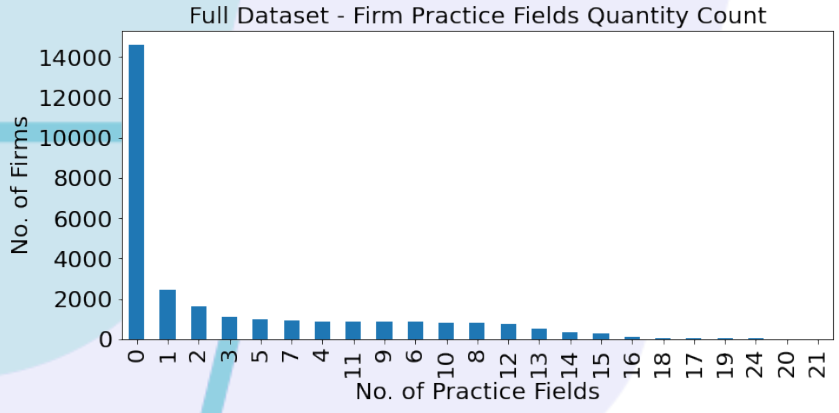
The final dataset contains 28823 rows, each row represents a LSP instance. The column information is summarised in the table below:

Column No.	Column Information	Description	Data Type
1	Authorized years	Feature 1	Integer
2	No. of Fields	Feature 2	Integer
3	Websites	Feature 3	categorical
4	Same Town	Feature 4	
5	No. of Offices	Optional Features (for future use)	Integer
6	Type		categorical
7	SSL encrypted		
8	Authorisation Type		
9	Authorisation Status		
10	Constitution		
11	Contacted	Target	

The following charts demonstrate the characteristics of the first 4 features of the dataset:

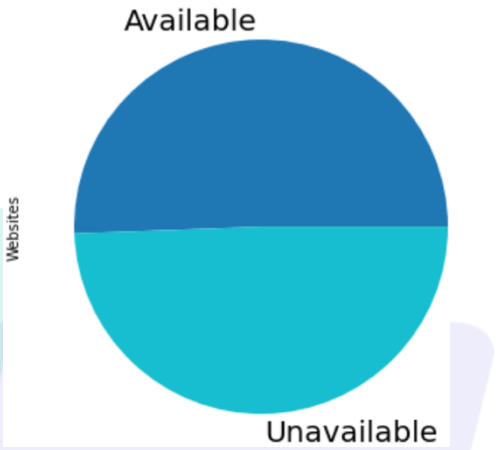


This chart depicts the number of years of SRA-authorisation vs the number of SRA-regulated entities. It implies that around 50% of entities have not accumulated a single years' authorisation from the SRA. It also implies that, of the remaining 50% with accumulated authorisation years, only 50% (6,000) have reached the 10+ years of authorisation history with the remaining 50% (6,000) combined having a near equal balance in numbers from 1 to 9-years' authorisation history.

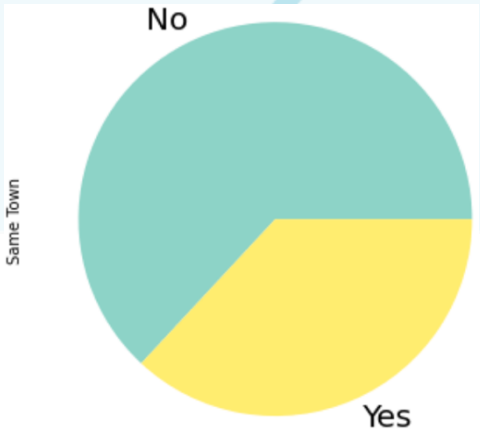


This chart depicts the number of "areas of practice" associated with each SRA-regulated entity. As provided in the chart, 14,000 of those entities had "0" areas of practice associated with it. The remaining firms are more distributed between 1-12 practice areas in which each entities covers.

However, from Legal Utopia's own review, this is representative only of the quality of the data held by the SRA and the inadequate annual collection of this datapoint, as the adopted labels for this datapoint are incomprehensibly vague for SRA entities to accurately associate a practice area with the services they provide, as well as those labels having significant overlap (Example: "'Tax Law", "Revenue Law", etc), regional application, and subjective interpretation that causes inaccuracies.



This chart depicts the balance of SRA entities with an available website address linked to their firm of all SRA entities. It provides for a 52%/48% availability in favour of "Available". This, however, does not factor website quality, security, loading, or display compatibility which, from Legal Utopia's own review, scores poorly across the 52% in all fields.



This chart depicts the proportion of SRA entities with more than one office or branch of all SRA entities. It implies that the vast majority (just under 75%) of SRA entities only have one office or branch.

3.8. Data Standardisation

The table below summarises the data type and value ranges of the feature columns within the dataset. The "Auth Year" and "No. of fields" range is much bigger than that of the "Websites" and the "Same Town" which just contain binary values. Using the original scale may cause biased weights on the "Auth Year" and "No. of fields" features and results in an inaccurate prediction. To resolve this problem, scaling technique will be applied to all columns of the dataset to fit all feature values into a uniformed range, so that the machine learning algorithm will not be biased toward the higher range features.[5]

[5] Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems, Chapter 4, Feature Engineering on Numeric Data

Feature No.	Column Information	Range	Data Type
1	Authorized years	0-20	Integer
2	No. of Fields	0-21	Integer
3	Websites	0 or 1	Binary
4	Same Town	0 or 1	

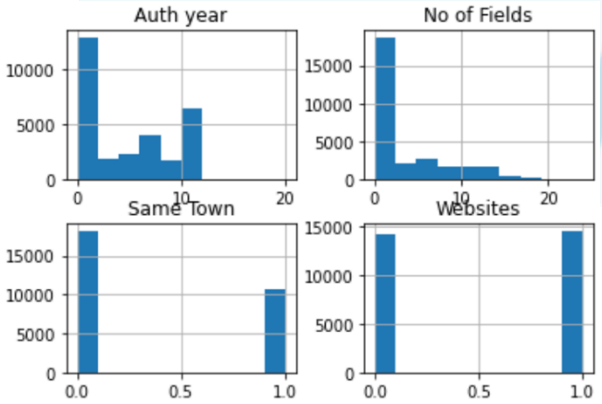
Scikit-Learn provides a data standardisation tool called "MinMaxScaler" which rescales features with a distribution value between 0 and 1. The minimum value of a feature is transformed into 0, and the maximum value is transformed into 1.[6] The general calculation formula is shown below:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

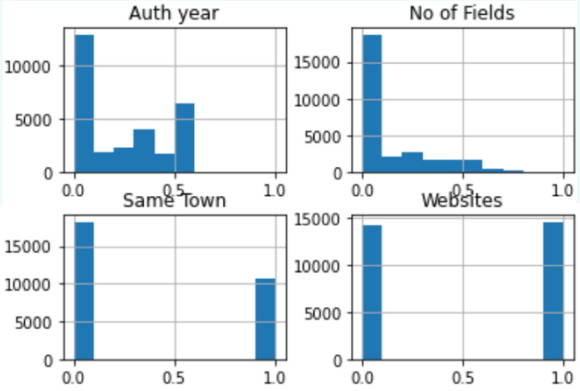
Where, x represents the original value, x^t represents the normalized value.

The histograms below show the feature value distribution prior and post data standardisation/scaling. It is noticeable that the data range is scaled to 0 to 1 uniformly but distribution shape is unchanged.

[6] Data Transformation: Standarization vs Normalization



Pre-Data Standardisation



Post-Data Standardisation

4. BUILDING MACHINE LEARNING MODELS TO PREDICT USERS' SELECTIONS

This section aims to build the machine learning models as proposed in Section 1.3 to predict user's selection on the full dataset prepared in Section 3.7.

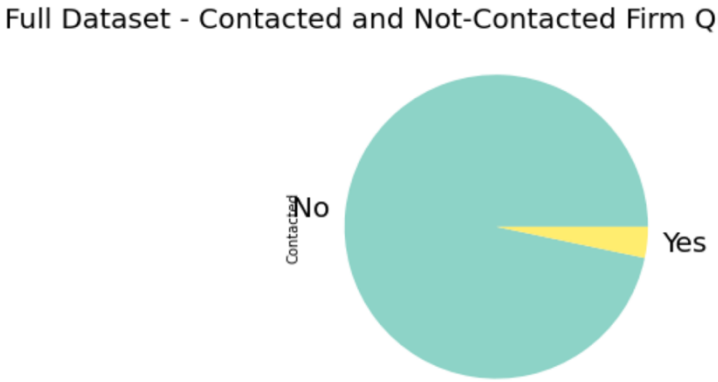
4.1. Dataset Check, Overview and Split

The dataset used for this section includes the first four features as listed in Table 3 and the target column. The full dataset contains 28823 rows x 5 columns.

It is split into training data (80%) and testing data (20%). The training data is used for model training and hyper-parameter turning validation purpose while testing data is reserved for prediction purposes only. The split sizes are summarised in the table below:

	Rows	Columns
Train features	23058	4
Train label	23058	1
Test features	5765	4
Test label	5765	1

A prominent problem of this dataset is the imbalanced distribution of the two classes. The contacted class only contains 951 samples while the not contacted class comprises 27872 samples. This is an extremely biased dataset and is suspected to be detrimental to the effect of machine learning. The plot below depicts the proportion of each class within the entire dataset.



4.2.
Logistic Regression Model

Linear machine learning models are considered a suitable algorithm for this project due to its simplicity and interpretability.

For this classification task, the Logistic Regression model holds advantage over the linear regression model as linear regression models treats the binary class labels 0 and 1 as ordinary integer numbers rather than categorical indication. Consequently, the

output of the linear regression model could go beyond 1 or lower than 0. It is hard to determine a meaningful threshold to distinguish the two classes.[7]

The Logistic Regression model is based on the sigmoid function as shown below, which ensures that the output is between the ranges between 0 and 1, which represents the probability range of 0 to 100%. The detailed prediction results will be discussed in Section 6.1.1.

$$P(Y=1 \mid \mathbf{X}) = \frac{1}{1+e^{-\mathbf{wx}}}$$

The Logistic Regression model is implemented using the API from the Scikit-Learn Library.[8] The hyper-parameters are turned using Stratified K-folds validation[9] and GridSearchCV tool[10] from the Scikit-Learn Library to suite the full dataset specifically, refer to the table below for details:

---

[7] Interpretable-machine-learning: A Guide for Making Black Box Models Explainable

[8] sklearn.linear_model.LogisticRegression

[9] sklearn.model_selection.StratifiedKFold

[10] sklearn.model_selection.GridSearchCV

Hyper Parameter	Parameter Function	Turning Options	Optimal Parameter
Penalty	specify the norm used in the penalization	'l1','l2','elasticnet'	l2
C	Inverse of regularization strength	20 samples between -4 to 4 on a log scale	78.47
Solver	Algorithm to use for optimization	newton-cg','lbfgs', 'liblinear', 'sag', 'saga'	saga

4.3. Support Vector Machine (SVM) Model

Support vector machine is one of the widely used algorithms for classification tasks. This algorithm tries to determine support vectors from the training data, which is turn establish a hyper-plane to separate the data points of each class in the hyper-space. A well-turned SVM maximises the distance between the hyper-plane and the closest data point of each class.

If the training data is linearly separable, the hyper-plane can be established straight forward, and this is called hard margin classification. Otherwise, the kernel functions are configured to map the data to higher dimensional space to enable the separation.[11] Different kernel options are tried at the hyper parameter turning stage.

[11] Machine Learning: A Probabilistic Perspective, chapter 14.5

The SVM model is implemented using the API from the SciKit-Learn Library[12] and the hyper-parameters are turned to suite the full dataset specifically. Refer to the table below for details:

Hyper Parameter	Parameter Function	Turning Options	Optimal Parameter
gamma	Kernel coefficient for kernel	1,0.1,0.01,0.001	1
C	Regularization parameter	0.1,1, 10, 100	0.1
kernel	Specifies the kernel type to be used	rbf', 'poly', 'sigmoid'	rbf

4.4 Nearest Neighbour (KNN) Model

The Nearest Neighbour (KNN) method is another widely used machine learning algorithm used for either classification or regression tasks. The way it works is straight forward, the algorithm finds K number of the nearest neighbours (according to the number of neighbours setting) for each data point, and count their class labels, then "draw" boundaries to separate different classes.

A key parameter of this algorithm is the number of neighbours. If it is set too small, the prediction is affected by outliers and create unstable decision boundaries. If it is set too big, it tends to class most data points to the majority class. The optimal value of this parameter will be cross validated at the hyper-parameter turning stage.

[12] sklearn.svm.SVC

A prominent disadvantage of the KNN model is the high computational expense. For each data point the algorithm calculate the distance to each training data sample. There are several approximation methods, and they will be cross validated at the hyper parameter turning stage.

The KNN model is implemented using the API from the Scikit-Learn Library[13] and the hyper-parameters are turned to suite the full dataset specifically. Refer to the table below for details:

Hyper Parameter	Parameter Function	Turning Options	Optimal Parameter
n_neighbors	Number of neighbors to use	integer from 1 to 30	28
p	Power parameter for the Minkowski metric	1, 2, 3, 4, 5	3
algorithm	Algorithm used to compute the nearest neighbors	ball_tree', 'kd_tree', 'brute'	ball_tree

4.5 Decision Tree Model

The term "Decision Tree" refers to a range of algorithms which share principles but are designed slightly different. The classification and regression trees (CART) is a popular decision tree algorithm and is likely adopted by Scikit-Learn for implementing a

[13] sklearn.neighbors.KNeighborsClassifier

decision tree API.[14] The CART algorithm finds the best features to split the dataset into subsets. It keeps doing so until the subset only contains samples from a single class or reached the maximum depth.[15]

The quality of each split is measured against impurity which can be configured to different options. This parameter is cross validated at the hyper parameter turning stage.

The "max_depth" parameter determines the maximum depth the algorithm can split dataset to. The deeper it is set to, the more information the algorithm can absorb from the data, and the longer time it takes to compute.

The "max_features" configure the number of features that are used as the splitting condition.

---

[14] The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark

[15] Machine Learning: A Probabilistic Perspective, chapter 16.2

The Decision Tree model is implemented using the API from the Scikit-Learn Library[16] and the hyper parameters are turned to suite the full dataset specifically. Refer to the table below for details:

Hyper Parameter	Parameter Function	Turning Options	Optimal Parameter
criterion	The function to measure the quality of a split	entropy','gini'	gini
max_depth	The maximum depth of the tree	integers from 1 to 32	2
max_features	The number of features to consider when looking for the best split	1, 2, 3, 4	3

4.6 Random Forest (RF) Model

Random Forest is a tree-based algorithm. Instead of employing a single tree, which is prone to over-fitting and variance error, Random Forest model constructs multiple trees and perform calculation in parallel. The result of each tree is aggregated at the end. In this way, the drawbacks of individual trees are weakened.

The Random Forest model is implemented using the API from the SciKit-Learn Library.[17] The number of trees is a key parameter; it determines the number of trees to ensemble. A higher number employs more trees at a cost of heavier computation load. This

[16] sklearn.tree.DecisionTreeClassifier

[17] sklearn.ensemble.RandomForestClassifier

setting can be configured at "n_estimators", this parameter will be cross validated at the turning stage.

The other hyper parameters as shown in the table below are all turned to suite the full dataset specifically. The parameters are application to each single tree in the same principle as the decision tree model.

Hyper Parameter	Parameter Function	Turning Options	Optimal Parameter
n_estimators	The number of trees in the forest	4, 8, 16, 32, 64, 100, 200, 400, 800, 1600	64
max_features	The number of features to consider when looking for the best split	auto', 'sqrt','log2'	log2
max_depths	The maximum depth of the tree	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110	4
min_samples_split	The minimum number of samples required to split an internal node	2, 5, 10	2
min_samples_leaf	The minimum number of samples required to be at a leaf node	1, 2, 4	4

4.7 Gradient Boosting Model (GBM)

The Gradient Boosting model is another type of tree-ensemble algorithm. Instead of building trees in parallel like the random forest model does, the gradient boosting model construct trees in series and the calculation is performed in a forward stage-wise manner.[18] This means the prediction result from a previous tree is passed onto the

[18] Decision Tree vs Random Forest vs Gradient Boosting Machines: Explained Simply

next tree, where the difference with the target is learnt and recalculated to reduce error. In this way, the results are "boosted" along the way till the last tree.

The Gradient Boosting model is implemented using the API from Scikit-Learn Library.[19] The hyper parameter to be cross validated are very similar to the Random Forest model, refer to the table below for details.

Hyper Parameter	Parameter Function	Turning Options	Optimal Parameter
n_estimators	The number of trees in the forest	50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750	700
max_features	The number of features to consider when looking for the best split	1, 2, 3, 4	1
max_depths	The maximum depth of the tree	integers from 1 to 32	2
min_samples_split	The minimum number of samples required to split an internal node	2, 5, 10	2
min_samples_leaf	The minimum number of samples required to be at a leaf node	1, 2, 4	1

4.8 XGBoost Model

The term XGBost refers to a gradient boosting library which provides efficient and flexible machine learning algorithms developed under the Gradient Boosting

[19] sklearn.ensemble.GradientBoostingClassifier

framework.[20] It shares a similar principle as the gradient boosting model developed by Scikit-Learn, the hyper parameter turning process is similar to. Refer to the table below for details.

Hyper Parameter	Parameter Function	Turning Options	Optimal Parameter
n_estimators	The number of trees in the forest	50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750	200
max_depth	The maximum depth of the tree	2,4,6,8	2

4.9 Multilayer Perceptron Neural Network

The last machine learning model implemented for this project is a Multilayer Perceptron Deep Neural Network. This model is constructed in a sequential manner using the Keras library. This model employs nine all connected dense layers and the neuron number of each layer are set in a gradual decreasing pattern, from 1024 to 4. The activation function of the dense layers is configured to "Relu", which is a linear function for positive inputs and faster to compute.

The last layer, which is the output layer using a sigmoid activation function to ensure the output is between 0 and 1 for the same reason as mentioned in section 4.2.

[20] XGBoost Documentation

This is binary classification problem, therefore the neuron number for the output layer is set to 1 and the loss function is set to "binary_crossentropy".

The "Adam" optimiser is a poplar stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments.***21***

4.10. Results Comparison and Discussion

All models have achieved a high accuracy figure 96% consistently. However, the overall accuracy figure does not reflect the whole truth. It is noticeable that no model recognised the samples from the contacted class. Both the precision and recall metrics are 0. This result suggests that no model learnt the patterns of the contacted class and predict all samples as not-contacted class. The overwhelming quantity of the not-contacted samples result in the overall high accuracy.

In fact, it is more critical to predict the contacted class samples, as the goal of this project is to recognise the characteristics of contacted LSPs, upon which their feature importance can be calculated.

---

[21] "Adam" optimizer.

It is suspected that the imbalanced data is the cause of the poor performance on the not-contacted class. Various experiments will be carried out in the next chapter to prove it and resolve the problem.

FULL DATA	Overall	Not contacted		Contacted	
	Accuracy	Precision	Recall	Precision	Recall
Logistic Regression Model	0.96	0.96	1	0	0
Support Vector Machine Model	0.96	0.96	1	0	0
Nearest Neighbour Model	0.96	0.96	1	0	0
Decision Tree Model	0.96	0.96	1	0	0
Random Forest Model	0.96	0.96	1	0	0
XGBoost Model	0.96	0.96	1	0	0
Gradient Boosting Model	0.96	0.96	1	0	0
Multilayer Perceptron Neural Network	0.96	0.96	1	0	0
Average	**0.96**	**0.96**	**1**	**0**	**0**

A confusion matrix is created for each model's prediction results. The decision tree model's confusion matrix is randomly picked for illustrating the algorithms' typical performance on both classes.

5. COST-SENSITIVE LEARNING FOR IMBALANCED CLASSIFICATION

This section aims to use cost-sensitive learning techniques to tackle data imbalance issues. Cost-sensitive learning is a subfield of machine learning related to classification on imbalanced data. Its goal is to minimise the prediction costs which are set

differently for different classes. Cost-sensitive learning includes multiple methods, including Cost-Sensitive Algorithms and Cost-Sensitive Resampling.[22]

Both methods are to be implemented in this section to compare their performance. The cost-sensitive resampling method includes over-sampling and under-sampling, different dataset variants will be generated accordingly. Machine learning models will be tested on the dataset variants to identify the most suitable dataset for this project. The hyper parameters of each machine learning model are turned to suite each specific dataset that they are trained on.

5.1. Cost-Sensitive Learning Algorithms

In general, machine learning models are trained on a dataset to minimise error or loss. When dealing with classification tasks, machine learning models assign equal weights to all classes by default. The Cost-sensitive Learning Algorithm idea is to set the cost for predicting the minority class incorrectly higher than that of the majority class when dealing with an imbalanced dataset. In this way, the machine learning model is expected to "focus" more on the minority class and achieve a satisfactory overall prediction results.

This is also the principle and motivation of its use.

[22] Cost-Sensitive Learning for Imbalanced Classification

5.1.2. Implementation

Scikit-Learn provides the means of adjusting the cost by configuring the "class_weight" argument, it will alter the cost related to misclassification of a certain class. The class weight is set in line with the cost related to misclassification of a certain class. The class weight is set in line with the quantity ratio of the contacted class and not contacted class, which is 951/27872 or 1/29/.1.

The following Scikit-Learn machine learning classifiers support the "class_weights" configuration:

- Logistic regression model
- SVM model
- Decision Tree model
- Random Forest Model

5.1.3 Data Preparation

There is no change to the dataset as used in Section 3.8.

5.1.4. Prediction Results Comparison

The table below summarised the machine learning algorithms prediction performance on the balanced dataset:

Cost-Sensitive Learning Algorithm	Overall	Not contacted		Contacted	
	Accuracy	Precision	Recall	Precision	Recall
Logistic Regression Model	0.96	0.96	1.00	0.00	0.00
Support Vector Machine Model	0.96	0.96	1.00	0.00	0.00
Decision Tree Model	0.96	0.96	1.00	0.00	0.00
Random Forest Model	0.96	0.96	1.00	0.00	0.00
Average	**0.96**	**0.96**	**1.00**	**0.00**	**0.00**

The prediction results are identical as the full data scenario as shown in section 0. It is noticeable that no model recognised the samples from the contacted class. Both the precision and recall figures are 0. This is an unsatisfactory result.

5.2. Prediction on a Balanced Dataset

5.2.1. Motivation

The cost-sensitive learning algorithm does not improve the prediction performance on the contacted class. From this section onwards, various cost-sensitive resampling methods will be implemented to resolve the imbalance issue. The most straight

forward approach is to under-sample the majority class data to match the quantity of the contacted class sample. With equal number of samples distributed, the machine learning algorithms performance is expected to improve, hence we can obtain a balanced dataset.

5.2.2. Generating a Balanced Dataset

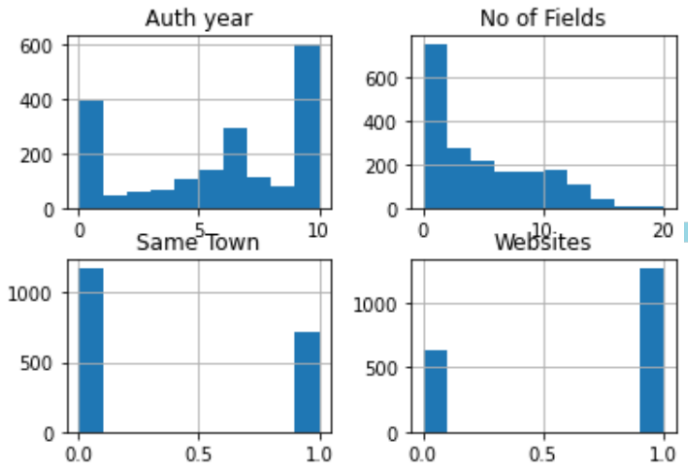
Generating a balanced databset is straight forward, retain all 951 contact class samples and random machine learning selection of 951 samples from the non-contacted class.

Label Distribution Count:

- Contacted: 951
- Non-Contacted: 951

The disadvantage of this down-sampling approach is that total data size is shrunk from 28823 to 1902. This is a dramatic size reduction, its effect on the prediction performance will be verified at the end.

Similar to Section 4, only the primary four features are retained for analysis, the overall balanced dataset size comes to 1902 rows x 5 columns. The features histograms are shown below:



The entire dataset is split into training data (80%) and testing data (20%). The training data is used for model training and hyper parameter turning validation purposes while the testing data is reserved for prediction purposes. The split data sizes are summarised in the table below:

	Rows	Columns
Train features	1521	4
Train label	1521	1
Test features	381	4
Test label	381	1

5.3. Prediction Results

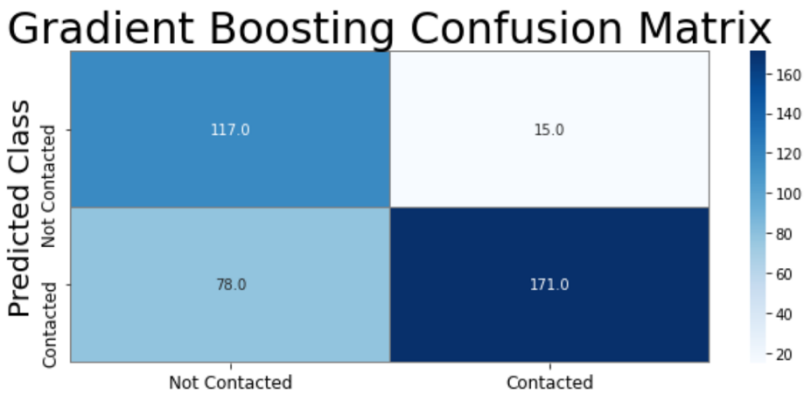
The table below summarises the machine learning algorithms prediction performance on the balanced dataset:

BALANCED DATA	Overall	Not contacted		Contacted	
	Accuracy	Precision	Recall	Precision	Recall
Logistic Regression Model	0.72	0.76	0.67	0.69	0.78
Support Vector Machine Model	0.68	0.81	0.5	0.62	0.88
Nearest Neighbour Model	0.75	0.87	0.61	0.69	0.9
Decision Tree Model	0.75	0.93	0.56	0.67	0.96
Random Forest Model	0.71	0.81	0.56	0.65	0.87
XGBoost Model	0.71	0.84	0.54	0.65	0.89
Gradient Boosting Model	0.76	0.89	0.6	0.69	0.92
Multilayer Perceptron Neural Network	0.74	0.88	0.57	0.67	0.91
Average	**0.73**	**0.85**	**0.58**	**0.67**	**0.89**

It is obvious that all models' performance on recognising the contacted class has improved dramatically. The recall metrics of the contact class is even much higher than that of the not contacted class. This result suggests that the imbalanced data was the cause of the poor performance on the contacted class in Section 3.8. The balanced dataset, although contains much fewer samples than the full dataset, improves all machine learning algorithms overall performance.

It is also noticed that all models' performance on the not-contacted class has decreased, particularly the recall figure. This is an indication of learning capacity decay on the not-contacted class.

A confusion matrix has been created for each model's prediction results. The Gradient Boosting model's confusion matrix is picked for illustrating the algorithms' typical prediction performance.



5.4. Prediction on a 2:1 Dataset

5.4.1. Motivation

Prediction on the balanced dataset has improved the contacted class precision and recall metrics dramatically. However, the performance on the not-contacted class has decreased. It seems the algorithms' learning focus has shifted towards the

contacted class too much. This section aims to resample the data to generate a dataset which contains two thirds of not-contacted class samples and one third contacted class samples.

5.4.2 Generating the Dataset

The dataset generation is straightforward, retain all 951 contact class samples and random machine learning selection of 1902 samples from the not-contacted class.

Label Distribution Count:

- Contacted: 951
- Uncontacted: 1902

Similar to Section 4, only the primary four feature are retained for analysis, the overall balanced dataset size comes to 2853 rows x 5 columns.

The entire dataset is split into training data (80%) and testing data (20%). The training data is used for model training and hyper parameter turning validation purposes while

the testing data is reserved for prediction purposes. The split data sizes are summarised in the table below:

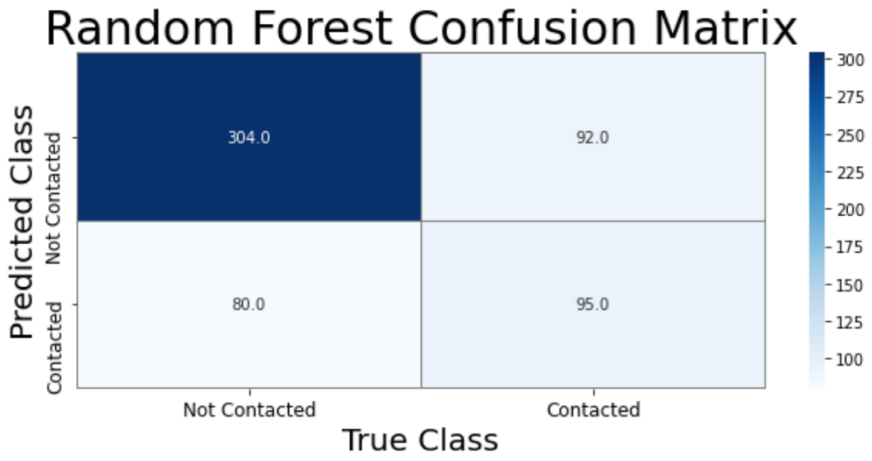
	Rows	Columns
Train features	2282	4
Train label	2282	1
Test features	571	4
Test label	571	1

5.4.3. Prediction Results

The table below summarises the machine learning algorithms prediction performance on the "2:1" dataset:

2:1 DATA	Overall Accuracy	Not contacted		Contacted	
		Precision	Recall	Precision	Recall
Logistic Regression Model	0.67	0.74	0.79	0.49	0.42
Support Vector Machine Model	0.67	0.71	0.88	0.5	0.26
Nearest Neighbour Model	0.7	0.77	0.79	0.55	0.52
Decision Tree Model	0.7	0.86	0.66	0.52	0.78
Random Forest Model	0.7	0.77	0.79	0.54	0.51
XGBoost Model	0.72	0.78	0.81	0.58	0.54
Gradient Boosting Model	0.71	0.8	0.77	0.55	0.6
Multilayer Perceptron Neural Network	0.67	0.67	1	0	0
Average	**0.69**	**0.76**	**0.81**	**0.47**	**0.45**

It is noticeable that the not-contacted class prediction performance has recovered to a certain degree, however, both the precision and recall metrics have dropped to an unacceptable level. Taking the confusion matrix of the neural network model for example (see below), only half of the contacted class sample are recognised (the right column of the matrix), which results in the poor recall metric of the contacted class. This result suggests that machine learning's performance is not linearly related to the number of samples of each class. Simply adjusting the sample number ratio is not successful approach for improving prediction performance.



5.5. Prediction on Synthetic Training Data

5.5.1. Motivation

A full but imbalanced data result is unsatisfactory whereas a balanced dataset has considerably improved overall performance, but not making full use of the available data (majority of the not-contacted class samples are not used). Less data could potentially lead to over-fitting and thus reduce the prediction accuracy.

This section attempts to generate a balanced data without sacrificing the overall data size. The prediction results will be compared with that of the balanced dataset to find out if a bigger data size will improve the prediction performance.

5.5.2 Synthetic Minority Oversampling Technique (SMOTE)

To maintain the original data size and achieve a balanced class sample distribution, the simplest solution is to duplicate the minority class samples to match the number of majority class samples. However, this approach just simply balanced the class distribution, but the duplicated samples do not provide any additional information.

An alternative solution is to generate synthetic samples based on the existing data. This idea is raised by Nitesh Chawla in 2002, and this method is called Synthetic Minority Oversampling Technique (SMOTE), which is a data augmentation technique.[23]

The way SMOTE works is that for each minority sample, it locates the 5 nearest neighbours (5 is the algorithm setting at the time the paper was released) from the same class. This depends on the oversampling ratio required, the algorithms randomly

[23] SMOTE: Synthetic Minority Over-sampling Technique

select a certain number of samples from the 5 nearest neighbours and draw lines between the neighbours and the underlying sample. Synthetic samples are then generated on those inter-connection lines.

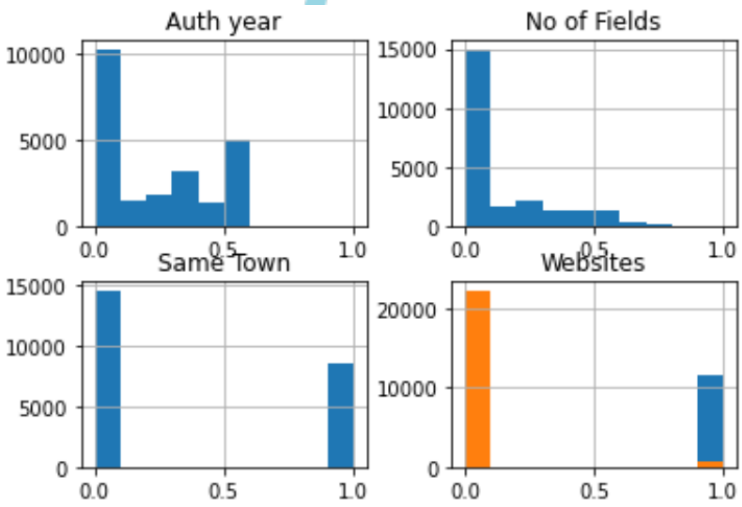
5.5.3 Generating a Balanced Training Dataset Using SMOTE

The SMOTE will only be applied for generating synthetic training data, so that a large number of balanced data will be made available for training machine learning models. The testing data, however, will only be reserved from the original dataset before applying SMOTE. Only in this way, the effect of the SMOTE method can be verified accurately.

The training data and testing data are firstly split in a 4:1 ration, as summarised below:

	Rows	Columns
Train features	44632	4
Train label	44632	1
Test features	5765	4
Test label	5765	1

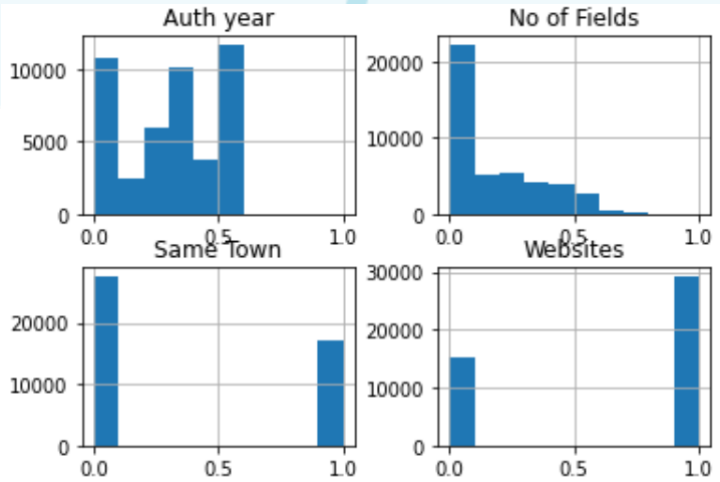
The histogram of the training data features before SMOTE are plotted for comparison purposes:



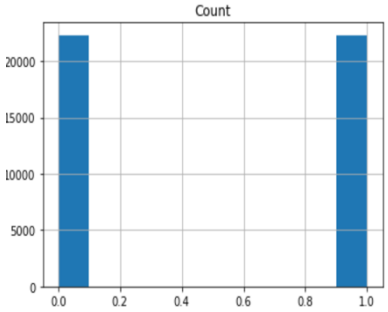
The imbalanced-learn Python library is used for generating synthetic training data. There is no need for setting any parameter, the program automatically generates synthetic samples for the minority class to match the quantity of the majority class. The code snippet is shown below:

```
X_train_oversample, y_train_oversample = oversample.fit_resample(X_train, y_train)
```

The histogram of the training features after applying SMOTE are plotted below. It is noticeable that the histogram shapes have changed, and the data size has increased.



The histogram of the training target after applying SMOTE is plotted below:



Training target Label Count after applying SMOTE:

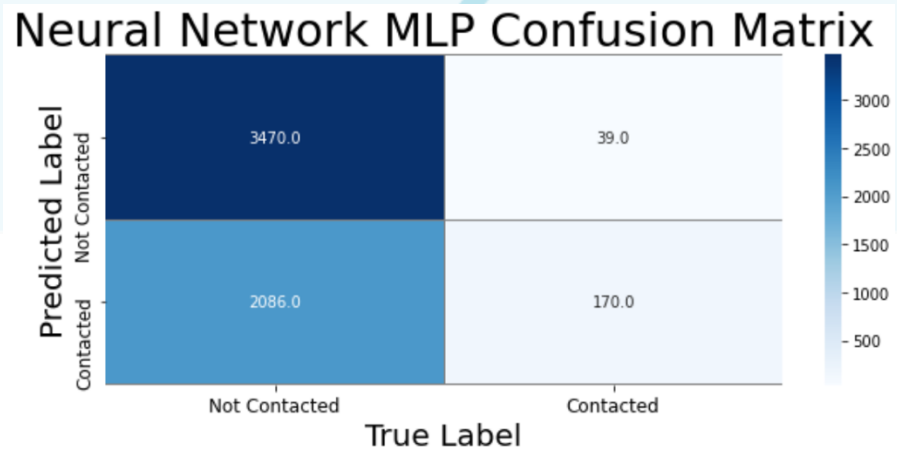
- Contacted: 22316
- Un-Contacted: 22316

5.5.4. Prediction Results

The table below summarises the machine learning algorithms prediction performance on the SMOTE dataset:

SMOTE DATA	Overall	Not contacted		Contacted	
	Accuracy	Precision	Recall	Precision	Recall
Logistic Regression Model	0.62	0.99	0.61	0.07	0.76
Support Vector Machine Model	0.51	0.99	0.49	0.06	0.93
Nearest Neighbour Model	0.71	0.98	0.72	0.08	0.61
Decision Tree Model	0.7	0.99	0.69	0.08	0.75
Random Forest Model	0.72	0.98	0.72	0.08	0.68
XGBoost Model	0.76	0.98	0.77	0.08	0.51
Gradient Boosting Model	0.76	0.98	0.77	0.08	0.49
Multilayer Perceptron Neural Network	0.63	0.99	0.62	0.08	0.81
Average	**0.68**	**0.99**	**0.67**	**0.08**	**0.69**

Comparing with the balanced data results in Section 5.2, the overall accuracy is lower. The biggest issue is extremely low contacted class precision metric obtained across all models. Taking the confusion matrix of the neural network model for example (shown below), most contacted class samples are recognised. However, a significant number of not-contacted samples are mistakenly recognised as contacted class (the left-bottom corner of the matrix), this is the cause of the extremely low contacted class precision metric.



5.6. Prediction on Balanced Dataset with more features

5.6.1. Motivation

The balanced dataset has achieved the best overall performance so far. This section attempts to include the additional features as listed in Section 3.7 for the classification. These optional features have not been displayed on the mobile app; hence users did not take them into consideration when making selections. However, these features reflect the LSP's quality and characteristics, the goal of this experiment is to discover if adding additional features would make any difference to the prediction results.

5.6.2. Generating the Dataset

The same principle as generating the balanced dataset is applied, the only different is the number of columns. The split data sizes are summarised in the table below:

	Rows	Columns
Train features	1521	9
Train label	1521	1
Test features	381	9
Test label	381	1

5.6.3. Prediction Results Comparison

The table below summarises the machine learning algorithms prediction performance on the SMOTE dataset:

BALANCED DATA WITH ADDITIONAL FEATURES	Overall Accuracy	Not contacted		Contacted	
		Precision	Recall	Precision	Recall
Logistic Regression Model	0.75	0.88	0.58	0.68	0.92
Support Vector Machine Model	0.76	0.96	0.55	0.67	0.98
Nearest Neighbour Model	0.73	0.79	0.65	0.69	0.82
Decision Tree Model	0.73	0.97	0.49	0.65	0.98
Random Forest Model	0.73	0.79	0.66	0.69	0.81
XGBoost Model	0.75	0.79	0.69	0.72	0.81
Gradient Boosting Model	0.75	0.78	0.69	0.71	0.8
Multilayer Perceptron Neural Network	0.72	0.8	0.61	0.67	0.84
Average	**0.74**	**0.85**	**0.62**	**0.69**	**0.87**

In comparison with the balanced data results in Section 5.2, the overall performance is almost identical, the differences in individual metrics are minimal. These results suggest that the additional features do not affect the machine learning algorithms' performance. This result is in line with the expectation as the additional features were not disclosed to users and hence did not take any effect in users' selection process.

5.7. Prediction Results Comparison Across all Dataset Variants

This section aims to compare the prediction results obtained from Section 5.1 to 5.5 and identify the most suitable dataset for the feature importance attribution used in Section 6.

The average metrics calculated in Section 5.1 to 5.5 and Section 3.8 are listed in the table for cross-comparison.

Note: The prediction results of the machine learning algorithms are stochastic to a certain degree, which means the model outcome may vary each time the prediction is repeated. This is more obvious for the deep neural network. For the balanced dataset, 2:1 dataset and the SMOTE dataset, random data sampling is required, the prediction result will vary as the data composition is different after re-sampling.

Cost-Sensitive/Dataset Variants	Overall	Not contacted		Contacted	
	Accuracy	Precision	Recall	Precision	Recall
Full Data	0.96	0.96	1.00	0.00	0.00
Cost-Sensitive Learning Algorithm	0.96	0.96	1.00	0.00	0.00
Balanced Data	0.73	0.85	0.58	0.67	0.89
Additional Features Balanced Data	0.74	0.85	0.62	0.69	0.87
2:1 Data	0.69	0.76	0.81	0.47	0.45
SMOTE Data	0.68	0.99	0.67	0.08	0.69

The full dataset and the cost-sensitive learning algorithms have the highest overall accuracy; however, it is achieved at the expense of giving up prediction of the contacted class. Therefore, these two methods/datasets are not considered the suitable candidate for the study in Section 6.

The Additional Features Balanced Dataset has won second place on the overall accuracy, followed by the Balanced Dataset. Their performance is almost indistinguishable. Bear in mind that the additional features have not been disclosed to users for their consideration but applied as filters by Legal Utopia of LSP characteristics that determine whether the LSP is included in the Find-A-Lawyer service at all. In contrast, the balanced dataset's features are all directly disclosed to users and, therefore, are more representative of the users' searching behaviour.

The 2:1 data and SMOTE data's overall accuracy metrics are the lowest among all candidates. More importantly, their performance on the contacted class prediction is unfavourable as compared with that of the balanced dataset. In fact, the contacted class prediction performance is the primary interest of this project, as the feature importance computation is based on the correct recognition of the contacted class samples. Due to their performance defect, these two datasets are not considered the suitable candidates for the study in Section 6.

Considering all factors listed above, the balanced data is deemed the most suitable dataset for the study carried in the next chapter where the feature importance analysis will be conducted.

6. *FEATURE IMPORTANCE COMPUTATION*

This section aims to use different feature attribution methods to compute the feature importance based on machine learning models' prediction on the balanced dataset. All methods' performance will be evaluated to identify the most suitable approach. At the end, the weights for the designated QIs will be calculated.

According to the requirement of this project, the feature importance method is required to intemperate the global feature importance. This section will use three commonly used approaches for computing feature weights respectively.

Note: All feature importance values are re-scaled in a way that the max feature value is scaled to 100% and all the other features' importance are compared with this max value and presented as a relative value. In this way, the feature importance values calculated by different methods can be cross compared.

6.1. *Native Feature Importance Interpretation*

Some of the interpretable machine learning algorithms that are implemented for this project are natively interpretable, which means their prediction results can be explained directly due to the algorithm's nature. These algorithms include linear regression, decision tree, random forest, XGBoost, and Gradient Boosting. The SVM, KNN, and neural network, however, do not support narrative features importance interpretation.[24]

6.1.1. Native Feature Importance of Linear Algorithm

The Linear model that is implemented for this project is the Logistic Regression, which provides the possibility for calculating the feature importance directly. Strictly speaking, the Logistic Regression model does not reflect a linear relationship between features and the target. However, its coefficients are proportional to the feature importance.

Normally, the logistic regression formula is given in the format as shown below:

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

[24] Interpretable Machine Learning- A Guide for Making Black Box Models Explainable, section 4

X are features and β are their weights. A simple transformation as shown below reflects the linear relationship.[25]

$$log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

The term in the log() function is called "odds". It represents the contacted class probability over the not-contacted class probability in the context of this project. "Odds" is equal to the sum of the features and their weights production. This equation can be fatherly transformed as:

$$\frac{P(y=1)}{1-P(y=1)} = odds = exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right)$$

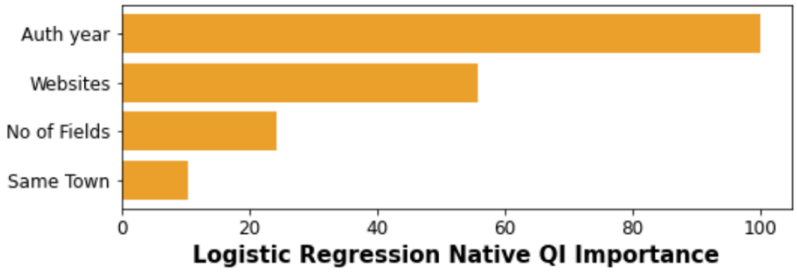
This formula suggests that a higher coefficient β contributes more toward "offs", and consequently, a higher probability of the contacted class. The amplitude of the coefficients provides the basis for a crude feature importance score.

The attribute "_coef_" provided by Sklearn can easily calculate the coefficients of the logistic regression model. The detailed results are listed in the below table:

[25] Interpretable Machine Learning- A Guide for Making Black Box Models Explainable, section 4.2

Features	Coefficients	Importance Ranking	Relative Weight (%)
Auth Year	1.82	1	100
No. of Fields	0.45	3	24
Same Town	0.19	4	10
Websites	1.02	2	56

The result is also plotted in the bar chart for visualisation:



There are some disadvantages associated with the logistic regression model:[26]

1. The accuracy of logistic regression models is not the highest among all machine learning algorithms used for these types of projects;

[26] Interpretable Machine Learning- A Guide for Making Black Box Models Explainable, section 4.4

2. Logistic regression can suffer from complete separation, i.e. if one of the features separates two classes very well, the weight for that feature will not converge. Fortunately, this does not occur for this project;

3. Logistic regression models struggle to handle the situations where the relationship between features and outcome is non-linear. This is not an obvious case for this project;

4. Logistic regression models is not the best option for the scenario that features interact with each other. The four features identified for this project are considered to be relatively independent as the logistic regression model's prediction accuracy is not far from the best performed algorithm.

6.1.2. Native Feature Importance of Tree-based Algorithm

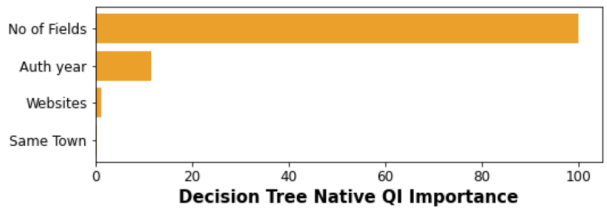
Half of the machine learning algorithms implemented for this project are tree based models, including decision tree, random forest, XGBoost and gradient boosting. Their mathematical principles are different, however, the way for calculating the native importance for those models are identical. These algorithms all provide a measure of feature importance based on the mean reduction in impurity.[27]

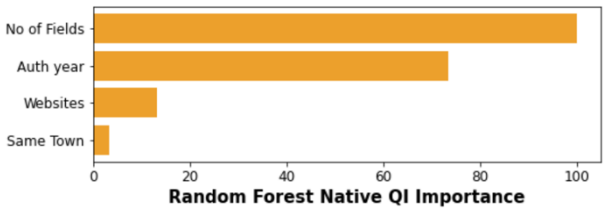
[27] Feature Importance Measures for Tree Models

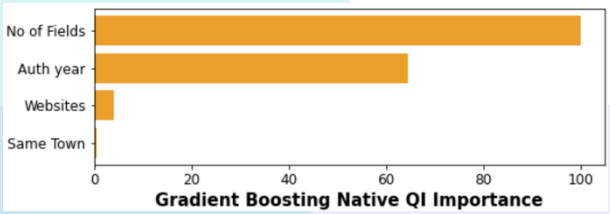
The attribute "feature_importances_" provided by Sklearn calculates the reduction in the "gini" criterion which is used to quantify the impurity.*282930* This is a native approach for calculating the feature importance for decision tree models. The detailed calculation results of each model are listed in the table below:

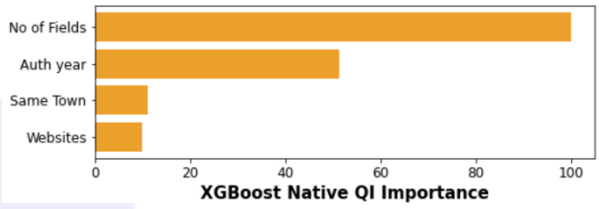
Native Importance	Auth Year		No. of Fields		Same Town		Websites	
	Ranking	Weight (%)	Ranking	Weight (%)	Ranking	Weight (%)	Ranking	Weight (%)
DT	2	12	1	100	4	0	3	1
RF	2	73	1	100	4	3	3	13
XGBoost	2	51	1	100	3	11	4	10
GBM	2	65	1	100	4	1	3	4

These results are plotted in the bar charts below for visualisation:









[28] sklearn.tree.DecisionTreeClassifier

[29] sklearn.ensemble.RandomForestClassifier

[30] sklearn.ensemble.GradientBoostingClassifier

It can be seen that the ranking results are relatively consistent across all algorithms, the only exception being the order of the "Same Town" and "Websites" features, which were swapped by the XGBoost algorithm.

It is noticeable that the weights of the "No. of Fields" and the "Auth Year" features overwhelm those of the other two features. This is because that impurity-based importance computation is biased towards high cardinality features.[31][32] In other words, if a feature contains a large number of unique values, it will gain high importance. The "No. of Fields" has the highest cardinality, and this caused the algorithm to think that it is the most important feature. The "Auth Year" feature has the second highest cardinality and, therefore, gained second place.

In contrast, the two binary category features, "Websites" and "Same Town", have the lowest cardinality, are believed to bring the least contribution towards the overall result by the tree-based algorithms.

---

[31] Explain your machine learning with feature importance

[32] Permutation Importance vs Random Forest Feature Importance (MDI)

6.1.3. Problems of the Native Importance Computation

There are several drawbacks associated with the native importance computation methods:

- The native importance values are based on training data statistics. Should the model be over-fitted, this approach may give high importance to features that are not predictive of the target in the unseen testing data.[33][34]
- This method relies on the machine learning type, algorithms like SVM, KNN, and deep neural network do not support native feature importance computation.
- Tree-type machine learning models may over-emphasise the features with high cardinality and under-estimate the contribution of the other features.

6.2. Permutation Importance Method

6.2.1. Permutation Importance Characteristics

---

[33] Interpretable Machine Learning- A Guide for Making Black Box Models Explainable, section 5.6

[34] Permutation Importance vs Random Forest Feature Importance (MDI)

To overcome the problems associated with native importance method, the Permutation Feature Importance approach is employed to calculate the importance in an alternative way.

The way Permutation Feature Importance is calculated based on the increase in the model's prediction error after permuting the feature values. The idea is that if a model's output relies on an important feature, shuffling this feature value will increase the model error. In contrast, shuffling an unimportant feature value, the model error will not change to a noticeable extent.[35]

In contrast to the native importance method, the Permutation Feature Importance approach is algorithm agnostic, it can be applied to all machine learning models developed for this project, enabling results comparison across all algorithms.

More importantly, permutation features importance is computed on unseen test data, in this way, misleading results are prevented in case the machine learning model is over-fitted at the training stage.[36]

---

[35] Interpretable Machine Learning- A Guide for Making Black Box Models Explainable, section 5.6

[36] Permutation feature importance

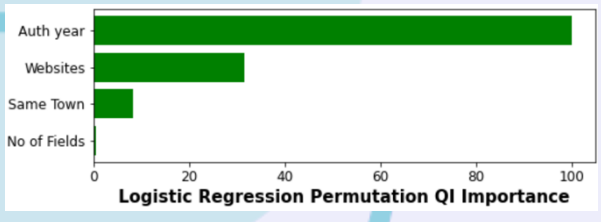
The Permutation Feature Importance method is generally considered as a relatively efficient technique that works well in practice. However, its disadvantage is that the importance of correlated features may be overestimated.[37]

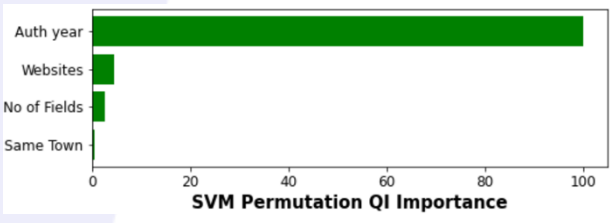
6.2.2 Feature Importance Results

The feature importance results are summarised in the table below:

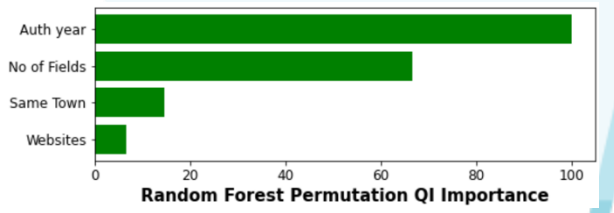
Permutation	Auth Year		No. of Fields		Same Town		Websites	
	Ranking	Weight (%)	Ranking	Weight (%)	Ranking	Weight (%)	Ranking	Weight (%)
LR	1	100	4	0	3	8	2	32
SVM	1	100	3	3	4	0	2	5
KNN	1	100	2	23	4	1	3	14
DT	3	0	1	100	4	0	2	0
RF	1	100	2	67	3	15	4	7
XGBoost	1	100	2	71	4	2	3	5
GBM	2	89	1	100	4	2	3	14
DNN	1	100	3	34	4	4	2	14

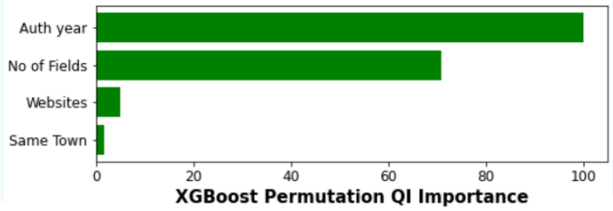
These results are plotted in the bar charts for visualisation:

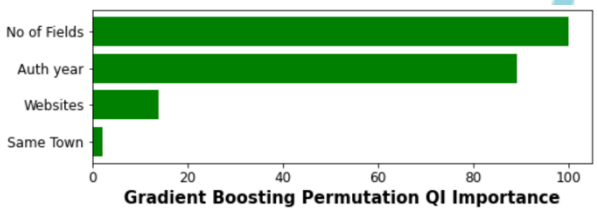


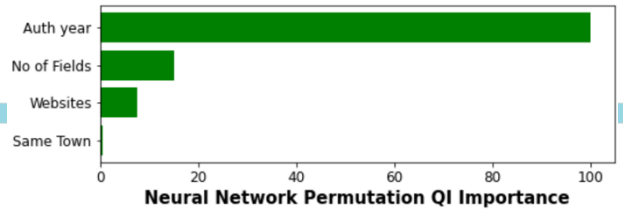


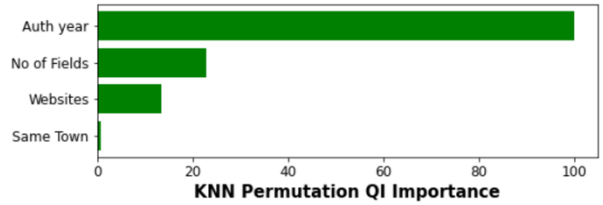
[37] Feature Importance Permutation

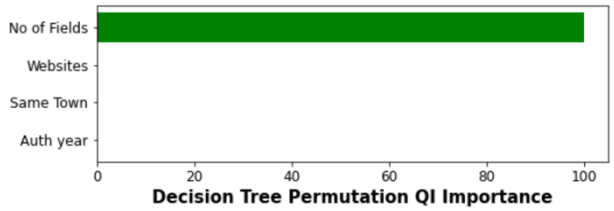












It is noticed that the feature importance ranking across all models is not as consistent as the native importance. The majority of models vote the "Auth Year" as the most influential feature and the "Same Town" as the least contributed feature. However, more diverse results are observed for the second and third places. The decision tree model has provided an extreme explanation, i.e. the "No. of Fields" dominates the model outputs while the other features do not contribute at all. The three ensemble models, namely random forest, XGBoost, and GBM assigned high importance weights to the two high cardinality features, "Auth Year" and "No. of Fields", which is in line with the observation in Section 6.1.2.

The results inconsistency observed above is believed to be associated with the mathematical principles of different algorithms. It is expected to test on another global interpretation method and compare its results with the permutation method.

6.3. SHAP Technique

This section attempts to utilise a popular machine learning interpretation approach, called SHAP, short for Shapley Additive explanations to compute the feature importance for the designated KPIs. The SHAP method is introduced by Lundberg and Lee in their paper "A Unified Approach to Interpreting Model Predictions". This approach is based on game theory, aiming to explain the output of any machine learning models.[38]

The SHAP technique provides both the global and local interpretabilities. The global interpretation refers to the features importance computation of the entire dataset, which is the goal of this project. The local interpretation targets at the individual data instances, providing interpretations on specific cases.

[38] A Unified Approach to Interpreting Model Predictions

6.3.1. How SHAP Works

In the game theory, all players participate into a game and contribute toward the game output. Similarly, in the SHAP concept, model output prediction is viewed as a game and all data features are viewed as the game players.

The SHAP approach firstly computes the marginal contribution brought in by each feature. Take this project for example, to work out the marginal contribution of the "Websites" feature, the SHAP method calculates the prediction difference with and without the "Websites" feature in any possible feature coalitions/combinations. There are four features in the dataset, the number of possible feature coalitions is 4!, which results in 24 different permutations.

The next step is to calculate the Shapley value of a feature as per the formula below:[39]

$$\varphi_i(v) = \frac{1}{n!} \sum_{R} \left[v(P_i^R \cup \{i\}) - v(P_i^R) \right]$$

Where, φ represents the Shapley value,

N represents the number of features,

i represents the "Websites" feature,

P_i^R represents the coalitions/combinations of the other 3 features

[39] A Unified Approach to Interpreting Model Predictions

$v(P_i^R)$ represents the contribution of the coalitions of the other 3 features,

$v(P_i^R \cup \{i\})$ represents the contribution of the "Websites" and the coalitions of the other 3 features,

The equation above implies that the Shapley value of a feature value is the average change in the prediction that the coalition already generates when the feature value joins them.[40]

The way that SHAP assign feature importance is according to the absolute Shapley value. To calculate the global feature importance, the absolute Shapley values of each feature are summed together across the entire dataset as per the equation below:[41]

$I_i = \sum_{j=1}^{n} |\varphi_i^j|$

Where, i represents the target feature,

j represents a data instance,

n represents the number of data rows,

φ represents the Shapley value of that feature

---

[40] Interpretable Machine Learning- A Guide for Making Black Box Models Explainable, section 5.10

[41] Interpretable Machine Learning- A Guide for Making Black Box Models Explainable, section 5.11

The bottleneck of Shapley value calculations is computation demand. For an N feature model, SHAP needs to compute 2^N distinctive models to calculate the Shapley value for a feature of a single instance. Therefore, sampling approximation is used to reduce the computation load. There are two approximation methods, one is Kernel SHAP which is model agnostic. The other is model specific approximation, developed for individual machine learning models. The model specific approximation is faster than Kernel SHAP.[42] The SHAP explainers and the corresponding approximation methods used for this project are summarised in the table below.

Note: There is no model specific explainer available for SVM and KNN models.

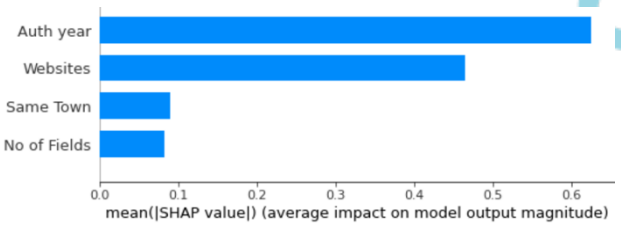
Algorithm	Explainer	Approximation Type
LR	LinearExplainer	Model Specific
SVM	KernelExplainer	Kernel SHAP
KNN	KernelExplainer	Kernel SHAP
DT	TreeExplainer	Model Specific
RF	TreeExplainer	Model Specific
XGBoost	TreeExplainer	Model Specific
GBM	TreeExplainer	Model Specific
DNN	DeepExplainer	Model Specific

6.3.2. Feature Importance Results

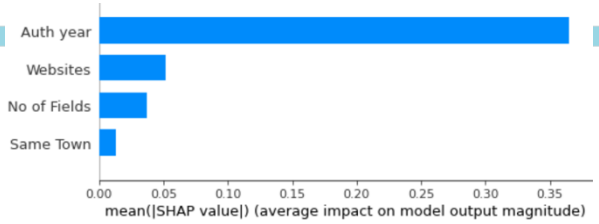
The feature importance results are summarised in the table below:

[42] A Unified Approach to Interpreting Model Predictions

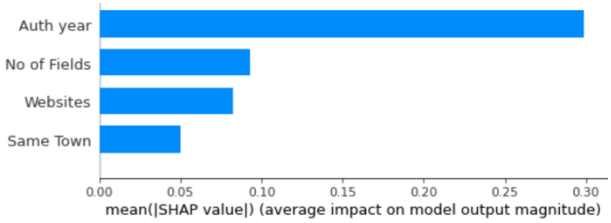
SHAP	Auth Year		No. of Fields		Same Town		Websites	
Importan	Ranking	Weight (%)	Ranking	Weight (%)	Ranking	Weight (%)	Ranking	Weight (%)
LR	1	100	4	13	3	15	2	74
SVM	1	100	3	10	4	4	2	14
KNN	1	100	2	31	4	17	3	28
DT	2	13	1	100	4	3	3	8
RF	2	90	1	100	4	13	3	28
XGBoost	1	100	2	47	4	6	3	22
GBC	1	100	2	58	4	7	3	20
DNN	1	100	3	19	4	12	2	60



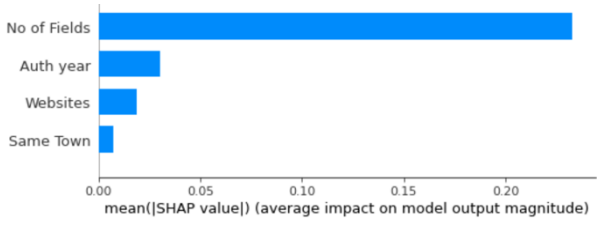
Logistic Regression SHAP



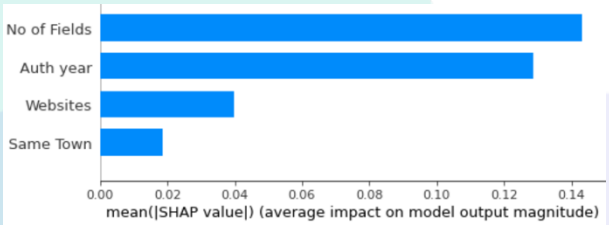
SVM SHAP



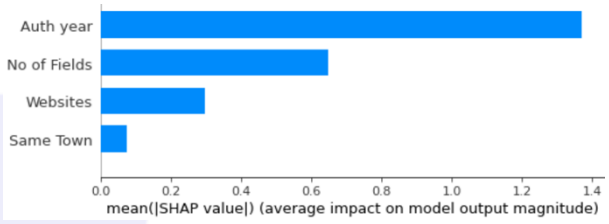
KNN SHAP



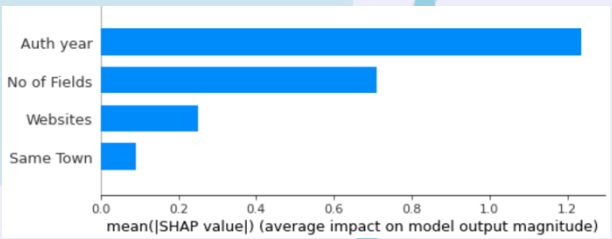
Decision Tree SHAP



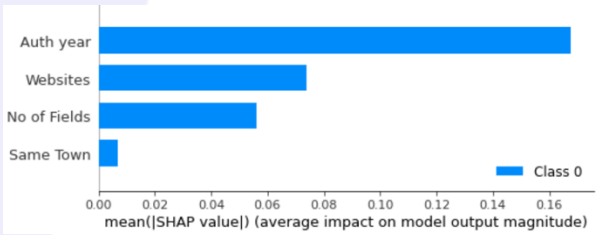
Random Forest SHAP



XGBoost SHAP



GBM SHAP



Deep Neural Network SHAP

Majority of the algorithms determine that the "Auth Year" is the most influential feature and the "Same Town" as the least contributing feature. The "Websites" has won more than half the votes for third place. However, the "No. of Fields" feature's ranking is the most controversial as it has won all possible rankings, from the first to the fourth.

6.4. Comparison of all Feature Importance Results

The table below summarises the feature importance results calculated by the three methods from Section 6.1 to 6.3 across all machine learning algorithms:

Interpretation	Algorithm	Auth Year		No. of Fields		Same Town		Websites	
		Ranking	Weight (%)	Ranking	Weight (%)	Ranking	Weight (%)	Ranking	Weight (%)
Native Importance	**LR**	1	100	3	24	4	10	2	56
	DT	2	12	1	100	4	0	3	1
	RF	2	73	1	100	4	3	3	13
	XGBoost	2	51	1	100	3	11	4	10
	GBC	2	65	1	100	4	1	3	4
Permutation Importance	**LR**	1	100	4	0	3	8	2	32
	SVM	1	100	3	3	4	0	2	5
	KNN	1	100	2	23	4	1	3	14
	DT	3	0	1	100	4	0	2	0
	RF	1	100	2	67	3	15	4	7
	XGBoost	1	100	2	71	4	2	3	5
	GBC	2	89	1	100	4	2	3	14
	DNN	1	100	3	34	4	4	2	14
SHAP Importance	**LR**	1	100	4	13	3	15	2	74
	SVM	1	100	3	10	4	4	2	14
	KNN	1	100	2	31	4	17	3	28
	DT	2	13	1	100	4	3	3	8
	RF	2	90	1	100	4	13	3	28
	XGBoost	1	100	2	47	4	6	3	22
	GBC	1	100	2	58	4	7	3	20
	DNN	1	100	3	19	4	12	2	60

It is noticeable that the feature importance results (both ranking and weight) obtained from three methods are not highly consistent. This is not uncommon as different interpretation methods are based on different mathematical principles and each type of machine learning model has its own characteristics. The fact is that there is no ground truth for interpreting/computing feature importance.[43] Except linear machine learning models, most machine learning models' inputs and outputs are not linearly related, and consequently a lack of intuitive method for attributing the feature importance.

6.5. Cross Evaluation of Feature Importance Attribution Methods

This section attempts to evaluate three feature attribution methods and identify the most suitable candidate for computing the QI weights.

The linear regression model provides high transparency for interpreting its predictions. The feature importance is directly reflected by its weights. This is the reason linear models have been widely used for interpretation-required applications or used as a

[43] Benchmarking Attribution Methods with Relative Feature Importance

surrogate model to explain "black-box" machine learning models.[44] The logistic regression model used for this project is not a purely linear model strictly speaking, however, it still provides good feature importance interpretability.

Unfortunately, either linear regression model or logistic regression perform well when the relationship between features and outcome is non-linear or, in the scenario where features interact with each other.[45] In this project, logistic regression model has achieved a medium performance. This is because the number of total features is not high, and the features are relatively independent from each other.

Due to the limitation of linear models, non-linear models, like tree-based or deep neural network models have been widely used. Various interpretation methods have been developed to intemperate non-linear algorithms. These interpretation methods are based on different mathematical principles, focusing on evaluating specific parameters/metrics and convert the measurement into features importance rations.

As pointed out in Section 6.1.3, the native importance reflects the interpretation of the training data and is subject to over-fitting issue. In contrast, the permutation and SHAP

---

[44] Interpretable-machine-learning A Guide for Making Black Box Models Explainable, Section 4.4

[45] Interpretable-machine-learning A Guide for Making Black Box Models Explainable, Section 4.4

methods compute the feature importance of the testing data. Hence the permutation and SHAP methods' performance is more convincing. The most suitable interpretation methods should be selected between these two.

Lundberg and Lee have compared SHAP with LIME and DeepLIFT interpretation methods, they proved that SHAP aligns with human intuition better and is the most consistent approach for explaining deep neural learning models among the three.[46] In 2019, Lundberg and Lee have conducted a detailed comparison between SHAP and the other popular tree ensemble interpretation methods including the permutation and Saabas methods are all inconsistent. In contrast, from the Shapley calculation formula, Lundberg inferred that SHAP has a solid theoretical foundation for providing consistent results suggests that the SHAP is the most accurate attribution method. Lundberg also emphasised that SHAP values are the most intuitive interpretation.

SHAP's superiority is reflected in the results obtained for this project. The table in Section 6.4 suggests that SHAP has the highest consistency of the "Auth Year" and the "Same Town" ranking results. The "Websites" ranking consistency is similar to the native importance method. The attribution of the "No. of Fields" features is inconsistent, but it is not worse than that of the permutation results.

[46] A Unified Approach to Interpreting Model Predictions

Futhermore, Christoph Molnar pointed to a key advantage of SHAP, which is the fastest computation of tree-based models. This feature makes a big difference when interpreting a large dataset. As discussed at Section 6.3.1, the SHAP method is based on Shapley value calculation. To obtain the global feature interpretation, which is the goal of this project, the Shapley values for the entire dataset shall be calculated and aggregated. Fast computation makes it practical and much more efficient to perform the enormous computation task.[47]

When calculating the feature importance for this project, it is noticed that the time consumed for permutation method is significantly longer than that of the SHAP model specific method.

Considering all factors, the SHAP method is deemed the most suitable approach for attributing the global feature importance for the designated QIs.

6.6. Computing the QI Weights

The final step is to apply the SHAP method to the best performed machine learning model. The table below is taken from Section 5.2 which summarises all algorithms performance on the balanced dataset.

---

[47] Interpretable-machine-learning A Guide for Making Black Box Models Explainable, Section 5.11.10

BALANCED DATA	Overall Accuracy	AUC	Not contacted		Contacted	
			Precision	Recall	Precision	Recall
Logistic Regression Model	0.72	0.75	0.76	0.67	0.69	0.78
Support Vector Machine Model	0.68	0.72	0.81	0.5	0.62	0.88
Nearest Neighbour Model	0.75	0.8	0.87	0.61	0.69	0.9
Decision Tree Model	0.75	0.77	0.93	0.56	0.67	0.96
Random Forest Model	0.71	0.79	0.81	0.56	0.65	0.87
XGBoost Model	0.71	0.79	0.84	0.54	0.65	0.89
Gradient Boosting Model	0.76	0.8	0.89	0.6	0.69	0.92
Multilayer Perceptron Neural Network	0.74	0.79	0.88	0.57	0.67	0.91
Average	**0.73**	**0.78**	**0.85**	**0.58**	**0.67**	**0.89**

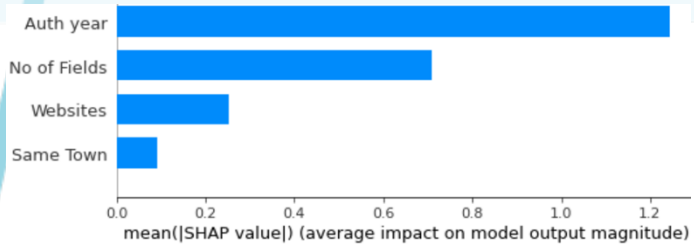
The Gradient Boosting model is considered as the most suitable algorithm for computing the feature importance. The reasons being:

- The highest overall accuracy
- The highest Area Under the Curve (AUC)
- The highest Contacted class recall, which is the primary interest of this project, as the feature importance computation is based on the correct recognition of the contacted class sample.

Hence, the finalised feature importance is calculated by applying SHAP method on the Gradient Boosting model. The results are listed below and plotted in bar chart for visualisation.

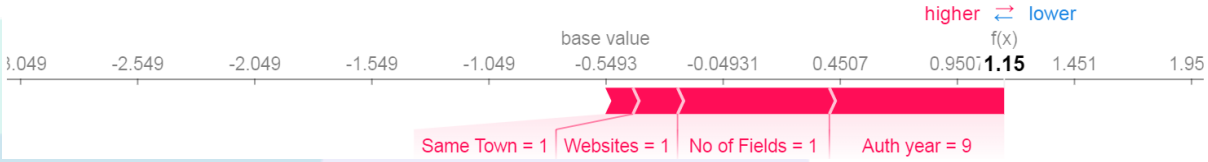
These weights shall be applied to calculate the overall score for each LSP.

Gradient Boosting	SHAP Importance	
	Rank	Weight (%)
Auth Year	**1**	100
No. of Fields	**2**	58
Same Town	**3**	20
Websites	**4**	7



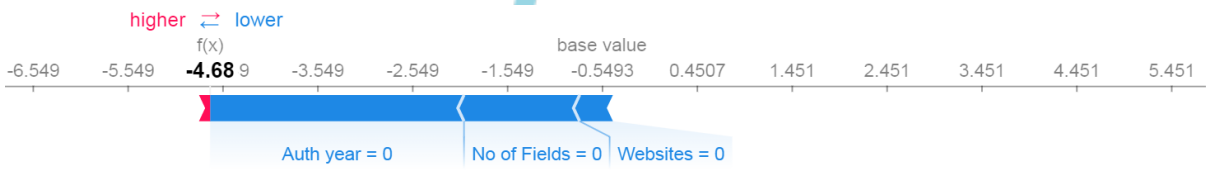
6.7 SHAP Explanation on the Gradient Boosting Model Outputs

This section uses the SHAP method to explain the Gradient Boosting prediction results to provide more detailed explanation at both global and individual levels.

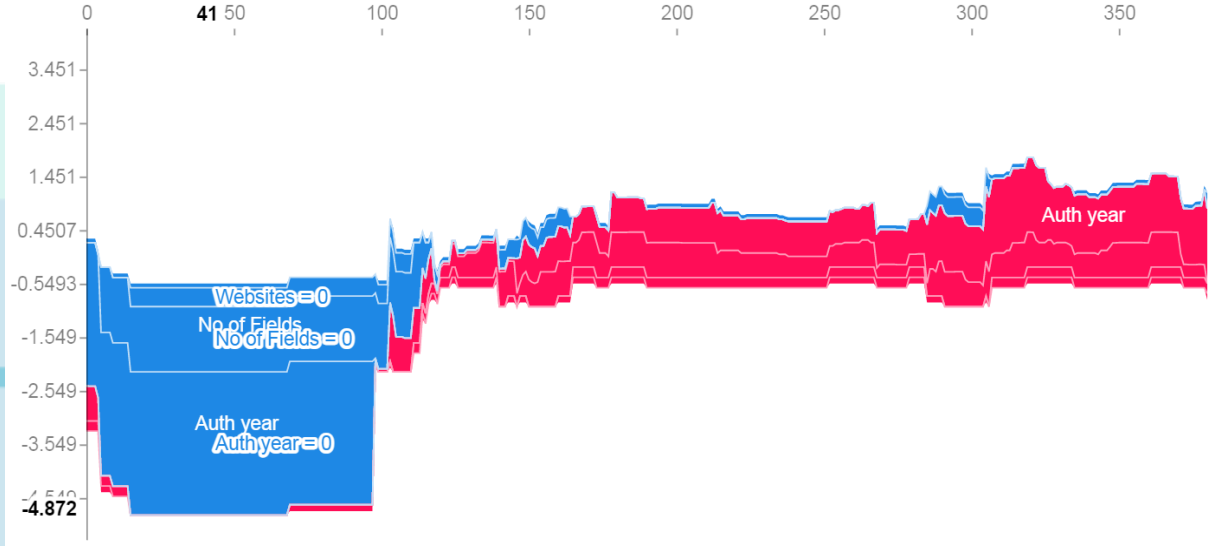


The plot above interprets the 1st sample in the testing data. The plot suggests that the "Auth Year" (9) and "No. of fields" (1) are the most "positive" influential features. They are followed by the "Websites" (1/available) and the "Same town" (1/Yes) features. The overall contribution of the four features pushed the model much higher than the base value (the average model output over the training dataset), hence this is a

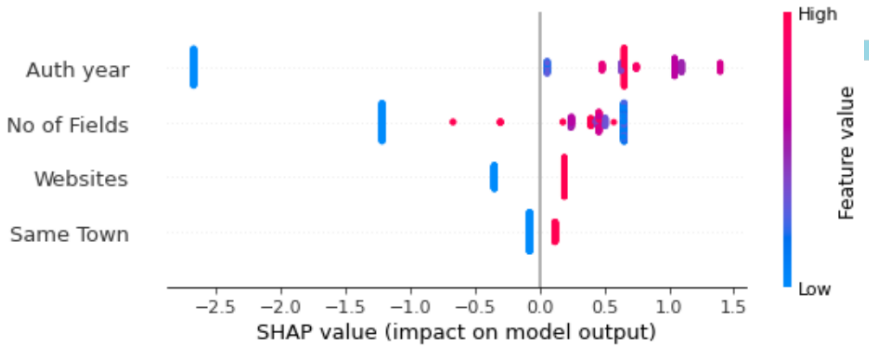
contacted case. The overall contribution of the 4 features pushed the model output much higher than the base value. This prediction is correct.



The plot above interprets the 3rd sample in the testing data. The plot suggests that the "Auth Year" (0) and "No. of fields" (0) are the most "negative" influential features. They are followed by the "Websites" (0/unavailable) and the "Same town" (0/No) features. The overall contribution of the 4 features pushed the model output much lower than the base value (the average model output over the training dataset) hence this is a not-contacted case. This class prediction is correct.



The plot above provides an overview of all testing data samples which are grouped as per their classes. The first 120 samples represented in the blue colour likely belong to the not-contacted class as their features values are mostly 0. The remaining testing samples are inclined toward the contacted class as most of their features have positive or high values.

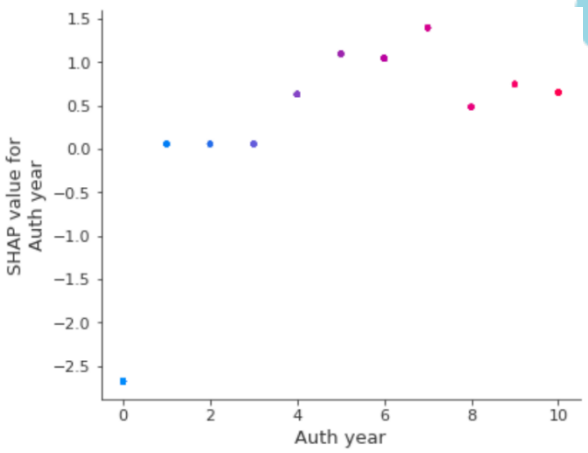


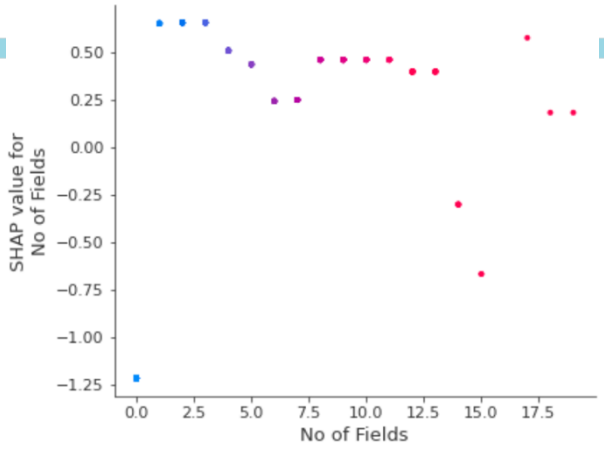
The plot above ranks the features importance from the highest on the top to the lowest at the bottom. A higher value of "Auth Year" has a higher feature value and impacts more on the model output, i.e., push the sample towards the contacted class, vice versa.

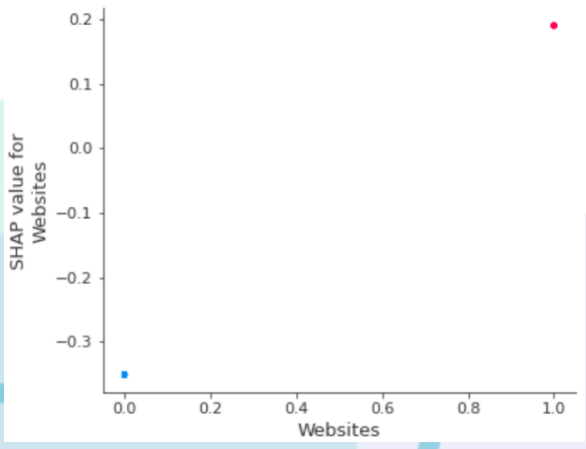
The "No. of fields" impact is non-linear as either a high or low value could push the sample towards the not-contacted class. This is in line with the observation made in Section 6.3.2.

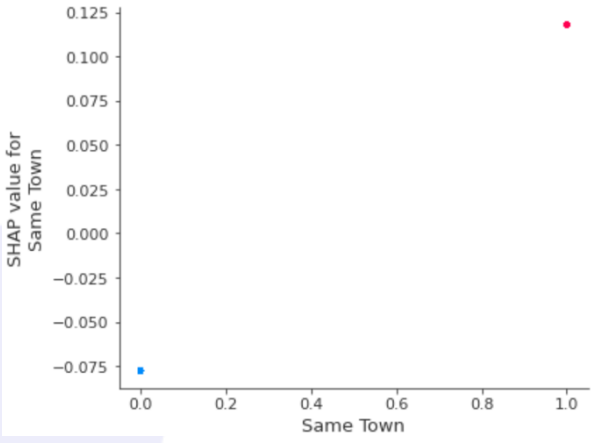
Both the "Websites" and "Same Town" feature are binary values. Value 1 push the sample class towards the contacted class while value 0 imposes the opposite effect.

Both of their overall impact magnitude (low absolute SHAP value) is smaller than “Auth Year: and “No. of fields”.









The four plots above depicted the feature’s importance characteristics individually.

The following observations are made:

- The "Auth Year" between 0 to 3 cause users to not contact the LSP while the value from 3 and above are likely to lead users to contact the LSP.

- For the "No. of fields", users tend to not contact LSPs if this value is 0 or 15. For the other values, it is likely to cause users to contact the LSP. This result is not intuitively logical, this observation aligns with the diverse rankings it obtained from all machine learning models.

- The interpretation for the "Websites" and the "Same Town" is straight forward, i.e. value 1 increases the user contact probability whereas value 0 decreases the probability.

7. CONCULSION

This project has attempted to analyse the Legal Utopia mobile app user data from the Find-A-Lawyer service to discover how the quality indicators influence a user's decision when selecting an LSP. To meet this objective, the following tasks have been performed and the key achievements are summarised below.

7.1 Dataset

A single dataset has firstly been generated to reflect customer's selection behaviour and all LSPs' characteristics including quality indicators and logistic information. This is achieved by organically merging the limited volume of mobile app user data and the limited number and accuracy of datapoints (features) available from the Solicitors Regulation Authority datasets.

Unnecessary LSP branches or offices identified and removed, so that the remaining entities approximately represent the LSPs that have been browsed but not selected by users. This information is not included in the mobile app data.

7.2 Modelling

Both classical machine learning and deep learning neural network algorithms have been implemented for recognising the commonalities/characteristics of the LSP that have been selected by customers and those not selected. The key hyper parameters of each machine learning model have been turned to suite a specific dataset.

7.3 Results

All machine learning algorithms have performed unsatisfactorily due to the severely imbalanced dataset. Several techniques have been employed to resolve the

imbalanced issues including cost-sensitive learning algorithms and cost-sensitive learning resampling.

The cost-sensitive learning algorithm method adjusts the incorrect prediction costs differently for the minority class. The costs difference is set in line with class size ratio. This approach was applied to the machine learning models which supports this functionality, however, the prediction results remain unsatisfactory.

The cost-sensitive learning resampling approach has been implemented by generating different dataset variants. The main focus was on the "balanced dataset" and the "SMOTE dataset". Both variants have re-balanced the sample numbers from both classes. The difference is that the "balanced dataset" abandoned a great number of samples from the majority class whereas the SMOTE method generates synthetic samples for the minority class. The experiment results suggest that a "balanced dataset" achieved the best overall performance and is used for the feature importance analysis.

Three methods have been used for calculating the feature importance based on the balanced dataset. The native feature importance method has been applied to the interpretable models including linear and tree-based models. The feature ranking consistency across these models is relatively good, however, the prominent problem of this method is that it reflects the feature importance of the training data. Since no machine learning model has achieved an extremely high accuracy due to the small

data, it is suspected that all models have been over-fitted to a certain degree, hence the interpretation on the training data is not reliable.

The main focus is shifted to the comparison of the permutation and SHAP methods performance. Both techniques are algorithm agnostic and interpret on the unseen testing data. By comparison, the SHAP method holds moderate advantage in terms of ranking consistency. This superiority could be more obvious should a larger dataset be used for analysis as the inventors of SHAP have conducted experiments to compare the performance of SHAP with various machine learning interpretation methods including the permutation method. The experiment results suggest that the SHAP method is superior in terms of both accuracy and consistency. Their conclusion also emphasises that the SHAP interpretation aligns with human intuition.

Finally, the SHAP method applied to the best performing Gradient Boosting model to calculate the feature importance value based on the balanced dataset. The computation results are presented in Section 6.6 along with detailed explanations in Section 6.7.

8. LIMITATIONS, SUGGESTIONS, AND FUTURE WORK

8.1. Limitations

The primary limitations associated with this project is the size and quality of the available data. At the time this project commenced, the mobile app user data contains 1140 samples, each sample represent a user searching event. Excluding the events where the user did not select any LSP, the useful sample number is 1044. In addition, the mobile app data lacks the records of LSPs that have been browsed but not selected. This missing information has to be substituted by the data from the SRA dataset approximately.

As a consequence of these two factors, the generated dataset for machine learning model training is severely imbalanced. Although several techniques have been used to cure this problem, the best performed model can only achieve a 76% overall accuracy and 80% AUC.

The feature importance interpretation is relying on the model prediction performance. Should a more accurate prediction be achieved, the importance ranking, and weights results are expected to be more consistent and representative across all models.

8.2 Suggestions

Based on the challenges encountered and observations made in the course of the project, the following suggestions are made aiming for developing a more accurate and consistent LSP scoring system in the future:

- The current mobile app only records the LSPs that have been contacted by users, lacking the records of the LSPs that have been browsed but not contacted. If the full user searching behaviour is recorded, machine learning algorithms will be able to identify the commonalities of both classes more accurately.

- Due to the relatively small number of the contacted class samples to the number of SRA entities, the entire dataset is severely imbalanced. Once more user data is collected, conducting this analysis again, it is expected to achieve a more accurate and consistent feature attribution result.

- The current mobile app displays excessive amounts of SRA entities on the Google Map interface. All drop pins have identical appearance and lack of signs for distinction between one another contributes to the lack of data features that can be used as QI features. The design could cause low search efficiency and randomise user selection. The suggestion is to add quality indication filters on the Find-A-Lawyer service, so that only LSPs that meet the users' quality expectations are displayed. This functionality will help users increase their awareness of the QIs and understand their impact on the LSP selection.

- Costs information is likely to be highly relevant to drop pin selection and could be displayed to the user via the Google Map interface to promote regional shopping by replacing the UI of the drop pin to the price point.

- The current mobile app only displays limited QIs on the user interface for users' information. This could be expanded to include direct reference to the trading duration (by reference to Companies House API data) or by authorisation duration using the authorisation data (by reference to the SRA API or BSB spreadsheet). Further QIs could include:

    - Cybersecurity Credentials
    - Customer Reviews
    - User Sentiment Feedback
    - Price / Cost Paid
    - Regulatory Compliance History
    - Number of authorised lawyers
    - Number of Payment Options
    - Insurance coverage (Cyber & Data, PII)

- This project has gained knowledge of the QI importance from the data perspective without assistance of domain knowledge. Should it be possible, invite professionals in the legal field and marketing specialists to rank QIs and

advise on their weights. Compare these results with the table in Section 6.4 to identify any differences or commonalities.

- If possible, add a function for users to provide sentiment feedback after contacting a selected LSP. This is a valuable feature for the machine learning algorithm to adjust QIs importance attribution.

- Offer the ability for users to filter by a range of QIs within the UI of the service, this will provide valuable data on the frequency of QIs selected and, therefore, implies the level of importance placed on them by the user, as well as important data of the relevance of different QIs by location of the LSP and user, as well as the type of legal domain knowledge sought.

8.3 Alternative Data – Use Case

As stated in the last section that the primary limitation associated with this project is the limited size of data. An alternative dataset has been found which could be used to test the effectiveness of the proposed approach of the proposed approach for this

project. This alternative dataset is provided by the travel agency Expedia and made available on the Kaggle website.[48]

The training data alone contains more than 300 million rows and 150 columns, covers customers' hotel searching behaviour in year 2013 and 2014. The hotel searching use case is deemed very similar to the LSP searching application. They both cover the user profile information, the searching target characteristics, and the selection events.

Although the alternative dataset will not help analysing the LSP QIs directly, it can be used to verify the performance of the machine learning models and interpretation methods proposed for this project. Should it be discovered that a large dataset can help improve the prediction accuracy and interpretation consistency, the proposed solution can be applied to the larger scale data collected by Legal Utopia in the future.

[48] Expedia Hotel Recommendations, Which hotel type will an Expedia customer book?

8.4 Find-A-Lawyer 2.0

In parallel to this study, Legal Utopia's developers have taken onboard feedback from this work to enhance the data collected, encourage greater searching behaviour of LSPs, and improve the user experience in order to achieve the ultimate objective of this study; to model and predict comparable LSPs for user comparison purposes, based on the same or similar search criteria, and computationally calculate the importance of various quality indicators that influence the behaviour of users to contact an LSP.

The latest version, released on the 16th August 2021, introduces several noteworthy additions to support the data framework needed to achieve this studies objective.

8.4.1. Information Slider

This firstly includes the introduction of a slider tab to compliment the Google Map interface located behind it. This provides the user with a preference to prioritise the view of the location of LSPs or the viewing of a refined number of LSPs.

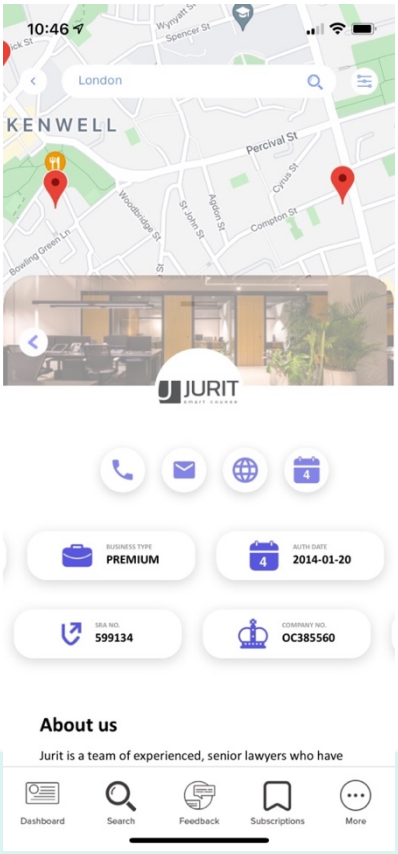
This slider is utilised as both a filter option to manage the introduction of different types of LSPs, such a government-contracted legal aid practitioners (LAPs) and public access barristers (PABs), to support greater browsing options.

It also introduces 'Standard' and 'Premium' member LSPs, currently exclusive to SRA-regulated entities, that are prioritised within the slider to increase user accessibility and reduce searching friction due to the overwhelming number of LSP options.

This will enhance the comparable information on filter preference of users based on LSP type, as well as the users' behaviour in selecting firms that are prominently displayed. Furthermore, the information available to the user is enhanced when visiting 'Premium' member LSPs with a number of new QIs that could influence user behaviour and selection, as well as the ability, for both tier member options, to book consultations with individual lawyers of the member LSP from the Find-A-Lawyer

service, this introduces a new option of engagement, as well as a significant number of new data features and metrics that could be used to achieve the study objective.

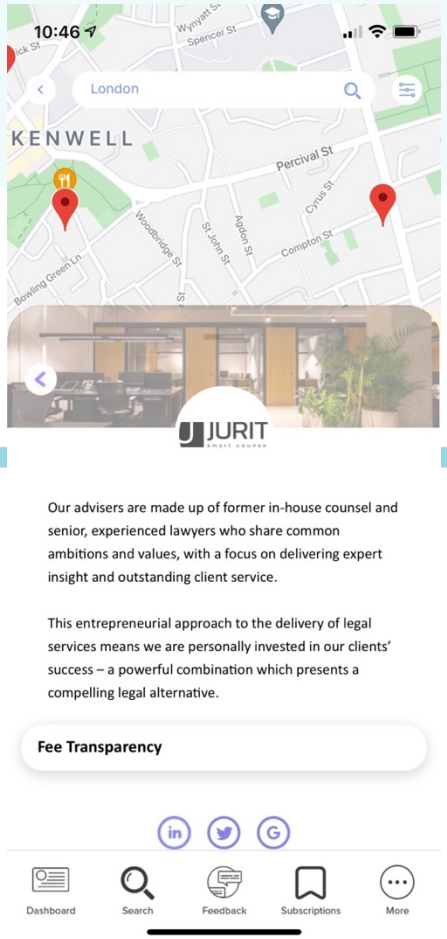
Premium Firm Profile



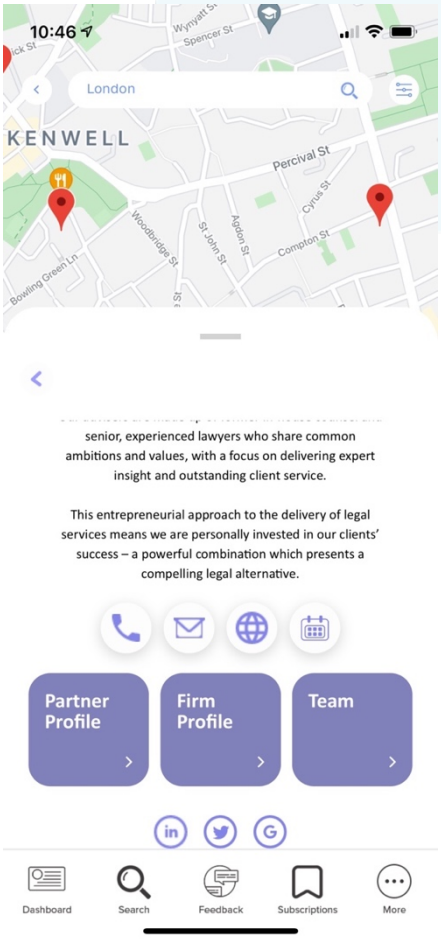
The premium firm membership was introduced for SRA-regulated law firms with a fee-earner team size of 10+ to market their firm, as well as their lawyers to users. It provided a means to import character and prominence, as well as the exclusivity of access to only users with a historical app-purchase or active subscription. This is with the view to increase the quality of the new client leads generated.

The profiles are more easily accessed on the Information Slider abovementioned to users, as selection from the Google Map interface remains overwhelming to most users, therefore, encouraging engagement to member firms.

The information displayed is in a unique format and design to other non-members and provides enhanced opportunities to bring prospective clients to contact the firm via their different channels of communication (social media, phone, email, etc).



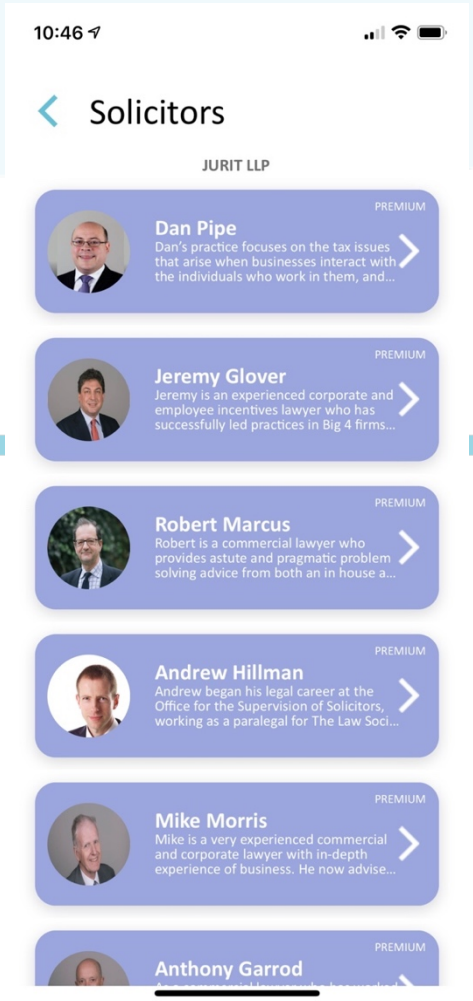
It also enables the firm to demonstrate its breadth of area of practice coverage with a team section comprising of all the lawyers listed on the platform. This has been identified as a key indicator that prospective clients seek as a reassuring component to contacting a firm.



There was also a need to facilitate access to background information that could be provided by the premium member firm themselves, this includes a dedicated partner profile. This profile allows, for instance, the Managing Director or Senior Partner to distinguish themselves as leading the firm.

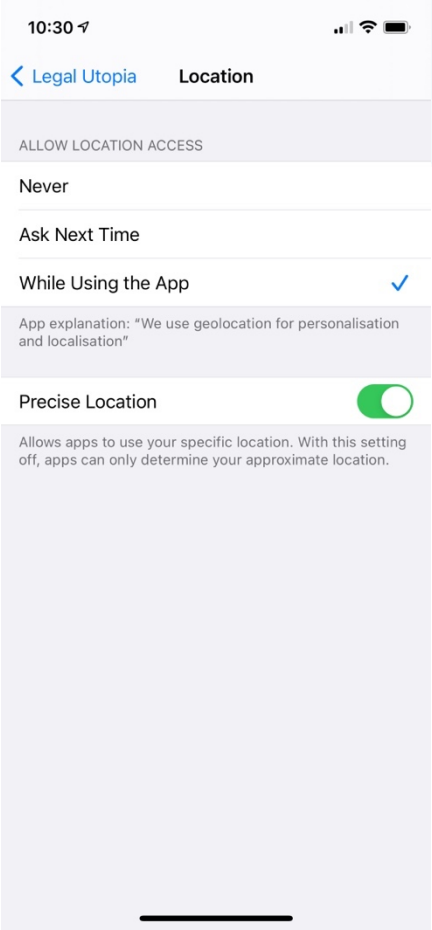
A key performance indicator for the premium firm member and Legal Utopia are the increased engagement analytics, the application enables the measurement of contact medium usage (no. times a call, email, web visit, or booking is made) from the premium profile.

A key benefit to a membership is the firm's connection between Find-A-Lawyer 2.0 and Book-A-Lawyer. This enables users to see the team of lawyers listed on the platform as a group, as well as review their profiles individually. However, the addition of the calendar button on the profile allows users to also view all lawyer profiles within Book-A-Lawyer to select from.



This searching and viewing behaviour can be monitored with the various datapoints collected to determine the increased likelihood of client engagement and bookings compared to non-members.

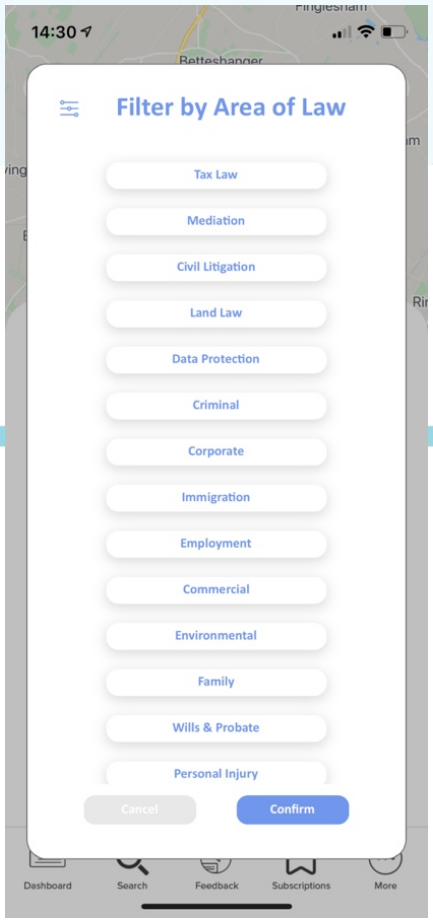
8.4.2. Geo-Location



The second feature update includes adding default geo-location within the background Google Map interface to automatically filter all LSPs based on the geo-location of the user's device. This is, on first use, only enabled subject to the user's consent.

This will provide greater accuracy in identifying if users seek LSPs based on their locality or whether they are willing to search further beyond a particular radius. This could also be measured based on the domain knowledge filters and the availability of legal domain knowledge in their locality to satisfy their legal needs.

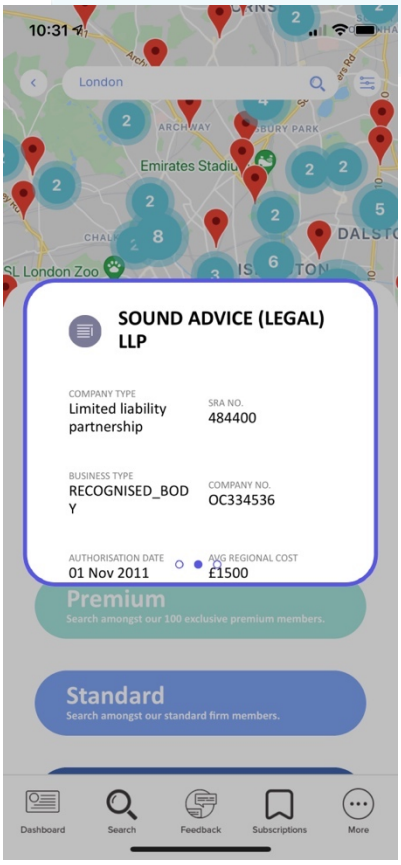
8.4.3. Legal Domain Filters



The third feature is not new but redesigned and more comprehensive, this is to update the areas of practice filter options to include a greater number of domains that are both standardised and more accurately applied to LSPs. This currently excludes BAPs due to a lack of service coverage information available from the Bar Standards Board.

The pre-existing SRA Organisations' data, as mentioned in Section 3.7, was inadequate to the point of having no reliable value for the purposes of its collection. As the data is not standardised, defined, comprehensive to each LSP, nor regularly updated (at the time of this study, this data is collected annually).

8.4.4. Firm Information & Costs



The SRA-regulated firms that are non-member LSPs now contain new "Regional Cost" information on every profile as a new QI feature which is believed to be influential to user searching behaviour. This "Regional Cost" is the median cost of a legal matter of firms in that firm's regional (example, Cambridgeshire, Manchester, London, etc) area, applied on a per branch/office level, after assessing the individual costs of several standardised legal matters generated by Legal Utopia from the raw pricing data of the Legal Services Board's prices dashboard in 2020.[49]

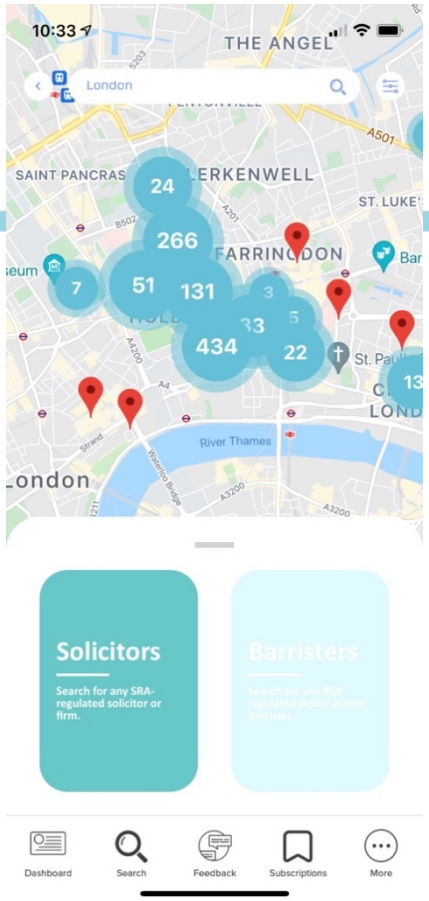
These matters are then linked to the areas of practice and geo-location of the LSP to ensure relevance (as such, the "Regional Cost" would apply based on (1) the LSP's location and (2) the LSP's area of practice coverage).

The SRA-regulated firms that are member LSPs now include the optional "Average Hourly Rate" information field on every profile as a new variable QI feature, which can be set or excluded by the LSP member, as well as the addition of a "Book Button"

[49] https://legalservicesboard.org.uk/price_research_dashboard_2020

which gives the user the option of viewing individual lawyer's information and the ability to book a video/audio consultation with the selected lawyer.

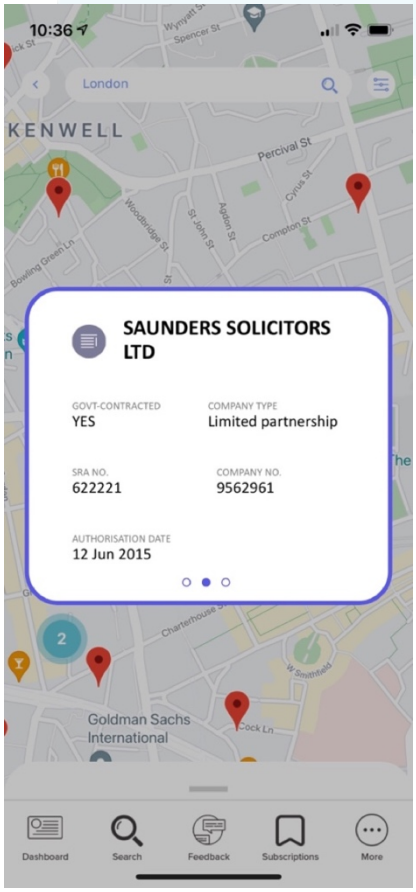
8.4.5. Public Access Barristers Information Tabs



The update to the Find-A-Lawyer service have introduced a new range of data points for Bar Standards Board regulated public access barristers, this was due to the relevance of the legal expertise that could be preferably accessed by users, as well as the comprehensive availability of the PABs' data; despite being unavailable as an API.

Due to the lack of absent information on the services provided by PABs, a filter or information tab display is unavailable. However, future work, in the absence of BSB engagement on this issue, will seek to research this missing element to create more comprehensive profiles.

8.4.6. Legal Aid Practitioners' Information Tabs



The introduction of more than 3,300 firms contracted by the Legal Aid Agency to provide legal aid funded work is also a new addition to the Find-A-Lawyer service. The filter option allows users to only display government-contracted firms on the map and then filter those LSPs by area of expertise in which the LSPs are authorised to provide legal aid work for.

Its introduction is early staged, but the later aim is to optimise the shopping experience for contracted legal aid practitioners following notification to users of the Legal Checker service of legal problem eligibility for civil legal aid, which is currently a free service across multiple areas of housing, debt, discrimination, etc.

The model produced in this study for comparison and quality scoring was not applied on the Legal Aid Agency directory of LSPs contacted to provide legal aid work, however, as the process to assessing and claiming civil legal aid becomes more digitised, it is envisaged that quality indicators could be separately developed to support comparison mapping and recommendation based on a variety of unique factors considered by those seeking legal aid practitioners. This early feature addition will allow Legal Utopia to collect initial data on searching and engagement with the service to inform this later work.

8.5 Stakeholder Engagement

A key inhibitor to this study was the lack of "live" data delivered through APIs from reliable and dependable sources, such as regulators, that collect and maintain the data. It is impractical to rely solely on market data that remains out-of-date, for instance, when a firm moves offices or closes a particular practice department.

In addition to the access to authoritative and reliable market data, there is a significant lack of foresight amongst existing datasets (live or not) that lack core principles of standardisation, both of datasets themselves and collaboration/communication between datasets. There is a disjointed ecosystem of data management and collaboration between government and industry funded data-holders, such as Solicitor Regulation Authority, Bar Standards Board, Legal Ombudsman, Chartered Institute of Legal Executives, and Council of Licensed Conveyancers.

8.5.1 Stakeholders

8.5.1.1 Solicitor Regulation Authority

Legal Utopia engaged with the SRA in October 2020 to provide a list of 24 additions or amendments to the SRA API following participation in a consultation on API use. Unfortunately, there was no further engagement or discussion on these points, nor any further changes between October 2020 and July 2021 on the data held as part of the SRA API or the functionality of the API.

On implementation of the SRA API by Legal Utopia, for the purposes of Find-A-Lawyer, the SRA API documentation (used by developers to implement programming) was incorrect. Legal Utopia engaged with the SRA to identify and correct this documentation, however, it indicated that other users of the SRA API were unlikely to be actively calling the API on a periodic basis to access up-to-date data.

It was further identified that the SRA API was established with only one query field, this means that the regulated firms listed within the dataset could only be called by name and not, for instance, postcode or regulatory status. This was a highly limiting factor to the API's use and would support the above presumption that other API users are not actively calling the SRA API on a periodic basis. This was raised with the SRA's technical staff with positive engagement on introducing additional query fields, however, no further response or changes have taken place between April 2021 and July 2021.

On implementation and data review (of the SRA Register API data), as of February 2021, there was no response to notification of subsequent identification of errors, missing, incomplete, or purportedly false data identified and no further update or engagement on the correction of this data when reported in May 2021.

8.5.1.2 QI Collective Regulator Consultation

In January and February 2021, via the LawTech UK Sandbox, Legal Utopia also engaged with the SRA, as well as the Bar Standards Board, Chartered Institute of Legal Executives, The Law Society of Scotland, Legal Services Board, and Legal Consumer Panel. A project scope and engagement materials concerning this study's outputs were also shared with these stakeholders, as well as forwarded onto the Council of Licensed Conveyancers.

Legal Utopia sought engagement on the following points:

1. What specific hurdles may there be on QI data points at a regulator level?
2. What specific hurdles may be envisaged on using individual or collective QIs on an entity basis?
3. What specific hurdles may there be on cross-regulator engagement?
4. What specific hurdles would you (at a regulator level) face in collecting, maintaining, sharing data at an entity level?
5. What is the specific hurdle you would face on engagement with this project?

6. What are your observations on the concept and cross-regulator proposal?

There was no engagement or response from The Law Society of Scotland or the Legal Consumer Panel. There was one collective email response from the SRA on behalf of the Bar Standards Board, Council of Licensed Conveyancers, and The Chartered Institute of Legal Executives, however, this engagement failed to meaningfully respond to the above engagement points. There was no further response or engagement to the above points since; despite reference to said further engagement, nor a response to subsequent queries.

In April 2021, Legal Utopia sought a portable copy of the SRA information database of regulatory history of SRA-regulated solicitors. This data was limited to only the regulatory history that was already in the public domain and published by the SRA on its own SRA Register. The request was made under the SRA's Transparency Code and was declined with Legal Utopia directed to the SRA Register for access on a solicitor-by-solicitor basis. This obstructed the ability to introduce regulatory history as an element of this study and an information filter and feature of Find-A-Lawyer 2.0.

The SRA later confirmed in May 2021 that said data was being used for a service as part of a joint collaboration/venture with other bodies. This would imply the availability of the data requested was, or could subsequently have been, made available to Legal Utopia.

8.5.1.3 Bar Standard Board

There was notification provided to the Bar Standards Board on the introduction of public access barristers to the Find-A-Lawyer 2.0 and an invitation to contribute to this in May 2021. There was no response or engagement to this invitation.

8.5.1.4 Chartered Institute of Legal Executives

There was positive engagement with the Charted Institute of Legal Executives on the production, and access to, a specific API to their register data of Legal Executives in January 2021 and a subsequent update in March 2021. Between January 2021 and July 2021, there has been no further consultation to the introduction of an API or access to it.

8.5.1.5 Companies House & Legal Services Board

The only meaningful and constant engagement on our engagement points above, including the assistant in the interpretation and access to research data, was from the Legal Services Board and Companies House. This contribution enabled elements of Find-A-Lawyer 2.0 to become feasible and subsequently launched to the market.

8.5.1.6 Parallel Output to Stakeholder Engagement

During the time engagement was sought with regulators and bodies in legal services, all but two engagement enquiry lines were considered successful. At the same time, Find-A-Lawyer, Book-A-Lawyer, and Find-A-Lawyer 2.0 were developed, tested, deployed, and upgraded as the largest sole repository of LSPs on a mobile application. This included a comprehensive review of all SRA and BSB-regulated entities, as of February 2021, establishing a more standardised, comprehensive, up-to-date database of regulated entities that those regulators.

8.6 Future Work

There is a significant amount of further work to be undertaken to assist the introduction of comparative analysis and quality indicators with a data-driven approach. This study has concluded that, although too early at this stage, with the right data balance between LSPs and historical user data, a comparison algorithm and QI importance computation could be feasible.

However, the study has indicated that there is a lack of a coherent and engaged data landscape amongst the specific regulators of regulated lawyers. The main barriers identified from this study are data access, data standardisation, and technical expertise amongst lawyer specific regulators.

A cross-jurisdictional data landscape will also be required to facilitate the incorporation and freedom of LSP shopping across the UK and Ireland. However, data

collection, sharing, and engagement are further behind outside of England and Wales.

In consideration of the above, this study sets out the following next steps to further a data-drive approach to LSP shopping, comparison recommendations, and quality indicator scoring:

1. To establish an open-source repository of the work undertaken, including the algorithms, programming, pre-processing, and datasets to facilitate industry engagement.

2. To establish key standardisation principles of key QI datapoints and communicate and lobby for the application of cross-regulator data standardisation.

3. To establish a data-trust of cross-regulator data following and in cooperation with points 1 and 2.

4. To establish a structured and comprehensive data framework and repository of historical user data for the purposes of furthering this study and establishing

the economic and commercial case for data-drive LSP comparison and QI importance weighting.

5. To hold regulators to account of their statutory objectives in the context of data collecting, management, and sharing as a result of the above.